N° d'ordre :

73-127

THESE

PRESENTEE

A L'UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

POUR OBTENIR

LE TITRE DE DOCTEUR - INGENIEUR

par

Serge CHAUDOURNE

ETUDE DU FILTRAGE NUMERIQUE RECURSIF DANS LE DOMAINE TEMPOREL

APPLICATION A LA SYNTHESE DES FILTRES ADAPTES

Soutenue le 29 juin 1973 devant la Commission d'Examen

MM.	R. PAUTHENET	Président
	J.L. LACOUME	Examinateurs
	J.F. CAVASSILAS	
	J. MAX	

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41534927_0048

N° d'ordre :

73-127

T H E S E

PRESENTEE

A L'UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

POUR OBTENIR

LE TITRE DE DOCTEUR - INGENIEUR

par

Serge CHAUDOURNE

ETUDE DU FILTRAGE NUMERIQUE RECURSIF DANS LE DOMAINE TEMPOREL

APPLICATION A LA SYNTHESE DES FILTRES ADAPTES

Soutenue le 29 juin 1973 devant la Commission d'Examen

MM. R. PAUTHENET	Président
J.L. LACOUME J.F. CAVASSILAS J. MAX	Examinateurs

UNIV.
LEIDEN
BIBL.

UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

LISTE DES PROFESSEURS

Président : Monsieur Michel SOUTIF
Vice-Président : Monsieur Gabriel Cau

PROFESSEURS TITULAIRES

MM.	ANGLES D'AURIAC Paul	Mécanique des fluides
	ARNAUD Georges	Clinique des maladies infectieuses
	ARNAUD Paul	Chimie
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOUD Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale
	BENOIT Jean	Radioélectricité
	BERNARD Alain	Mathématiques Pures
	BESSON Jean	Electrochimie
	BEZES Henri	Chirurgie générale
	BLAMBERT Maurice	Mathématiques Pures
	BOLLIET Louis	Informatique (IUT B)
	BONNET Georges	Electrotechnique
	BONNET Jean-Louis	Clinique ophtalmologique
	BONNET-EYMARD Joseph	Pathologie médicale
	BONNIER Etienne	Électrochimie Electrométallurgie
	BOUCHERLE André	Chimie et Toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques Appliquées
	BRAVARD Yves	Géographie
	BRISSONNEAU Pierre	Physique du Solide
	BUYLE-BODIN Maurice	Electronique
	CABANAC Jean	Pathologie chirurgicale
	CABANEL Guy	Clinique rhumatologique et hydrologie
	CALAS François	Anatomie
	CARRAZ Gilbert	Biologie animale et pharmacodynamie
	CAU Gabriel	Médecine légale et Toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques Pures
	CHARACHON Robert	Oto-Rhino-Laryngologie
	CHATEAU Robert	Thérapeutique
	CHENE Marcel	Chimie papetière
	COEUR André	Pharmacie chimique
	CONTAMIN ROBERT	Clinique gynécologique

MM.	COUDERC Pierre	Anatomie Pathologique
	CRAYA Antoine	Mécanique
Mme	DEBELMAS Anne-Marie	Matière médicale
MM.	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DESRE Pierre	Métallurgie
	DESSAUX Georges	Physiologie animale
	DODU Jacques	Mécanique appliquée
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	DUGOIS Pierre	Clinique de Dermatologie et Syphiligraphie
	FAU René	Clinique neuro-psychiatrique
	FELICI Noël	Electrostatique
	GAGNAIRE Didier	Chimie physique
	GALLISSOT François	Mathématiques Pures
	GALVANI Octave	Mathématiques Pures
	GASTINEL Noël	Analyse numérique
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques Pures
	GIRAUD Pierre	Géologie
	KLEIN Joseph	Mathématiques Pures
Mme	KOFLER Lucie	Botanique et Physiologie végétale
MM.	KOSZUL Jean-Louis	Mathématiques Pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques Appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre	Mathématiques Appliquées
	LEDRU Jean	Clinique médicale B
	LLIBOUTRY Louis	Géophysique
	LOUP Jean	Géographie
Mle	LUTZ Elisabeth	Mathématiques Pures
MM.	MALGRANGE Bernard	Mathématiques Pures
	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Séméiologie médicale
	MASSEPORT Jean	Géographie
	MAZARE Yves	Clinique médicale A
	MICHEL Robert	Minéralogie et Pétrographie
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NEEL Louis	Physique du Solide
	OZENDA Paul	Botanique
	PAUTHENET René	Electrotechnique

MM.	PAYAN Jean-Jacques	Mathématiques Pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET René	Servomécanismes
	PILLET Emile	Physique industrielle
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REULOS René	Physique industrielle
	RINALDI Renaud	Physique
	ROGER Jean	Clinique de pédiatrie et de puériculture
	SANTON Lucien	Mécanique
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SILBERT Robert	Mécanique des fluides
	SOUTIF Michel	Physique générale
	TANCHE Maurice	Physiologie
	TRAYNARD Philippe	Chimie générale
	VAILLAND François	Zoologie
	VALENTIN Jacques	Physique Nucléaire
	VAUQUOIS Bernard	Calcul électronique
Mme	VERAIN Alice	Pharmacie galénique
M.	VERAIN André	Physique
Mme	VEYRET Germaine	Géographie
MM.	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale
	YOCCOZ Jean	Physique nucléaire théorique

PROFESSEURS ASSOCIES

MM.	BULLEMER Bernhard	Physique
	HANO JUN-ICHI	Mathématiques Pures
	STEPHENS Michaël	Mathématiques Appliquées

PROFESSEURS SANS CHAIRE

MM.	BEAUDOING André	Pédiatrie
Mme	BERTRANDIAS Françoise	Mathématiques Pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques appliquées
	BIAREZ Jean-Pierre	Mécanique
	BONNETAIN Lucien	Chimie minérale
Mme	BONNIER Jane	Chimie générale
MM.	CARLIER Georges	Biologie végétale
	COHEN Joseph	Electrotechnique
	COUMES André	Radioélectricité
	DEPASSEL Roger	Mécanique des Fluides
	DESPORTES Charles	Chimie minérale
	GAUTHIER Yves	Sciences biologiques
	GAVEND Michel	Pharmacologie
	GERMAIN Jean-Pierre	Mécanique

MM.	GIDON Paul	Géologie et Minéralogie
	GLENAT René	Chimie organique
	HACQUES Gérard	Calcul numérique
	JANIN Bernard	Géographie
Mme	KAHANE Josette	Physique
MM.	MULLER Jean-Michel	Thérapeutique
	PERRIAUX Jean-Jacques	Géologie et minéralogie
	POULOUJADOFF Michel	Electrotechnique
	REBECQ Jacques	Biologie (CUS)
	REVOL Michel	Urologie
	REYMOND Jean-Charles	Chirurgie générale
	ROBERT André	Chimie papetière
	DE ROUGEMONT Jacques	Neurochirurgie
	SARRAZIN Roger	Anatomie et chirurgie
	SARROT-REYNAULD Jean	Géologie
	SIBILLE Robert	Construction Mécanique
	SIROT Louis	Chirurgie générale
Mme	SOUTIF Jeanne	Physique générale

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

Mle	AGHIUS-DELORD Claudine	Physique pharmaceutique
	ALARY Josette	Chimie analytique
MM.	AMBLARD Pierre	Dermatologie
	AMBROISE-THOMAS Pierre	Parasitologie
	ARMAND Yves	Chimie
	BEGUIN Claude	Chimie organique
	BELORIZKI Elie	Physique
	BENZAKEN Claude	Mathématiques Appliquées
	BILLET Jean	Géographie
	BLIMAN Samuel	Electronique (EIE)
	BLOCH Daniel	Electrotechnique
Mme	BOUCHE Liane	Mathématiques (CUS)
MM.	BOUCHET Yves	Anatomie
	BOUVARD Maurice	Mécanique des Fluides
	BRODEAU François	Mathématiques (IUT B)
	BRUGEL Lucien	Energétique
	BUISSON Roger	Physique
	BUTEL Jean	Orthopédie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHIAVERINA Jean	Biologie appliquée (EFP)
	CHIBON Pierre	Biologie animale
	COHEN-ADDAD Jean-Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie médicale
	CONTE René	Physique
	COULOMB Max	Radiologie

MM.	CROUZET Guy	Radiologie
	DURAND Francis	Métallurgie
	DUSSAUD René	Mathématiques (CUS)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	GENSAC Pierre	Botanique
	GIDON Maurice	Géologie
	GRIFFITHS Michaël	Mathématiques Appliquées
	GROULADE Joseph	Biochimie médicale
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et Médecine préventive
	IDELMAN Simon	Physiologie animale
	IVANES Marcel	Electricité
	JALBERT Pierre	Histologie
	JOLY Jean-René	Mathématiques Pures
	JOUBERT Jean-Claude	Physique du Solide
	JULLIEN Pierre	Mathématiques Pures
	KAHANE André	Physique générale
	KUHN Gérard	Physique
	LACOUME Jean-Louis	Physique
Mme	LAJZEROWICZ Jeannine	Physique
MM.	LANCIA Roland	Physique atomique
	LE JUNTER NOël	Electronique
	LEROY Philippe	Mathématiques
	LOISEAUX Jean-Marie	Physique Nucléaire
	LONGEQUEUE Jean-Pierre	Physique Nucléaire
	LUU DUC Cuong	Chimie Organique
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et Médecine préventive
	MARECHAL Jean	Mécanique
	MARTIN-BOUYER Michel	Chimie (CUS)
	MAYNARD Roger	Physique du Solide
	MICHOULIER Jean	Physique (I.U.T. "A")
	MICOUD Max	Maladies infectieuses
	MOREAU René	Hydraulique (INP)
	NEGRE Robert	Mécanique
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (IUT B)
	PEFFEN René	Métallurgie
	PELMONT Jean	Physiologie animale
	PERRET Jean	Neurologie
	PERRIN Louis	Pathologie expérimentale
	PFISTER Jean-Claude	Physique du Solide
	PHELIP Xavier	Rhumatologie
Mle	PIERY Yvette	Biologie animale
MM.	RACHAIL Michel	Médecine interne
	RACINET Claude	Gynécologie et obstétrique

MM.	RENAUD Maurice	Chimie
	RICHARD Lucien	Botanique
Mme	RINAUDO Marguerite	Chimie macromoléculaire
MM.	ROMIER Guy	Mathématiques (IUT B)
	SHOM Jean-Claude	Chimie Générale
	STIEGLITZ Paul	Anesthésiologie
	STOEBNER Pierre	Anatomie pathologique
	VAN CUTSEM Bernard	Mathématiques Appliquées
	VEILLON Gérard	Mathématiques Appliquées (INP)
	VIALON Pierre	Géologie
	VOOG Robert	Médecine interne
	VROUSSOS Constantin	Radiologie
	ZADWORNY François	Electronique

MAITRES DE CONFERENCES ASSOCIES

MM.	DOUDOURIS Georges	Radioélectricité
	CHEEKE John	Thermodynamique
	GOLDSCHMIDT Hubert	Mathématiques
	SIDNEY STUARD	Mathématiques Pures
	YACOUD Mahmoud	Médecine légale

CHARGES DE FONCTIONS DE MAITRES DE CONFERENCES

Mme	BERIEL Hélène	Physiologie
Mme	RENAUDET Jacqueline	Microbiologie

Fait le 30 mai 1972

Cette étude a été effectuée au Centre d'Etudes Nucléaires de Grenoble.

Avant de faire la présentation de ces travaux, je me dois de remercier tout d'abord Monsieur le Professeur R. PAUTHENET, Directeur de l'Ecole Nationale Supérieure d'Electrotechnique et de Génie Physique, d'avoir bien voulu me faire l'honneur de présider ce jury.

Je désire également exprimer toute ma reconnaissance à Monsieur le Professeur J.L. LACOUME qui a dirigé mes travaux, pour ses conseils et son aide.

Je tiens à remercier Monsieur le Docteur J.F. CAVASSILAS, du Centre d'Etude des Phénomènes Aléatoires de Grenoble, d'avoir accepté de faire partie du jury.

J'adresse à Monsieur CORDELLE, Directeur du L.E.T.I. toute ma gratitude pour m'avoir accueilli dans son laboratoire, et avoir mis à ma disposition les moyens matériels nécessaires.

Je tiens à remercier Monsieur J. MAX, Chef de la Division du Traitement du Signal, qui a eu l'idée de cette étude et qui m'a fait bénéficier pendant son élaboration de sa grande expérience et de ses conseils. C'est en outre son enseignement qui m'a permis d'acquérir les connaissances théoriques nécessaires.

Enfin, je remercie tous mes camarades de la Division Traitement du Signal auprès desquels j'ai toujours trouvé les renseignements, techniques ou théoriques dont j'avais besoin. C'est leur sympathie constante qui a rendu agréable ma présence au L.E.T.I.

TABLE DES MATIERES

REMERCIEMENTS

TABLE DES MATIERES

NOTATIONS ET SYMBOLES UTILISES

INTRODUCTION 9

Chapitre 1 - FORMALISME ET DEFINITION DES FILTRES NUMERIQUES 11

I - RAPPEL DES PROPRIETES DE L'ESPACE BL2 11

I.1 - Définition de BL2 11

I.2 - Structure et propriétés de BL2 12

I.2.1 - Propriétés des signaux de BL2

I.2.2 - BL2 est un espace de Hilbert

I.2.3 - BL2 est stable pour l'opération de convolution

I.3 - La base temporelle de BL2 14

II - DEFINITION DES FILTRES NUMERIQUES 15

II.1 - Les transformations linéaires dans BL2 15

II.1.1 - Définition

II.1.2 - Transformations bornées

II.1.3 - Convergence des transformations linéaires dans BL2

II.2 - Les filtres linéaires dans BL2 16

II.2.1 - Les translations

II.2.2 - Les filtres linéaires

II.2.3 - Théorème fondamental des filtres linéaires

II.3 - Les filtres numériques dans BL2 17

II.3.1 - Définition

II.3.2 - Convergence des filtres numériques

II.3.3 - Expression du filtrage dans la base temporelle de BL2

II.3.4 - Exemple de filtres numériques

II.3.4.1 - Filtre numérique identité

II.3.4.2 - Filtres numériques de translation

III - FILTRES NUMERIQUES PHYSIQUEMENT REALISABLES 19

III.1 - Les signaux physiques 19

III.2 - Les signaux dans les systèmes digitaux 20

III.3 - La causalité pour les filtres numériques 20

III.4 - L'espace BL2 des filtres numériques et des signaux physiques réels. 21

Chapitre 2 - LES FILTRES NUMERIQUES RECURSIFS 23

I - EQUATION AUX DIFFERENCES DEFINISSANT UN FILTRE NUMERIQUE RECURSIF 23

I.1 - Cas des systèmes continus 23

I.2 - Suite numérique solution d'une équation aux différences finies 23

I.2.1 - L'application

I.2.2 - Cas des suites à termes nuls pour k<o(ensemble S_+)

I.2.3 - Commutativité de $\mathcal{H}$ avec les translations dans S_+

I.2.4 - Expression directe de l'application $\mathcal{H}$

I.2.5 - Caractéristique de l'application $\mathcal{H}$

I.3 - Définition des filtres numériques récursifs 26

II - INTERETS TECHNOLOGIQUE DES FILTRES NUMERIQUES RECURSIFS 26

III - DETERMINATION DE LA REPONSE IMPULSIONNELLE A PARTIR DES COEFFICIENTS. CONDITION DE STABILITE 27

III.1 - Equation de convolution définissant un filtre numérique récursif 27

III.2 - Détermination de la réponse impulsionnelle à partir de la fonction de Green 28

III.3 - Calcul de la fonction de Green d'un filtre numérique récursif 28

III.4 - Expression de la réponse impulsionnelle. Condition de stabilité 29

III.4.1 - Expression de la réponse impulsionnelle

III.4.2 - Condition de stabilité

IV - METHODES D'ETUDE DES FILTRES NUMERIQUES RECURSIFS DANS LE DOMAINE FREQUENTIEL 31

IV.1 - Gain complexe d'un filtre numérique récursif 31

IV.1.1 - Expression du gain complexe

IV.1.2 - Condition pour qu'une expression du type (20) représente le gain complexe d'un filtre numérique récursif

IV.1.3 - Remarque sur le comportement à l'infini de la réponse impulsionnelle d'un filtre numérique récursif

IV.2 - Fonction de transfert en Z d'un filtre numérique récursif 32

IV.2.1 - Expression de la fonction de transfert en Z

IV.2.2 - Condition de stabilité dans le plan des Z.

IV.3 - Calcul de la norme d'un filtre numérique récursif 32

IV.3.1 - Définition

IV.3.2 - Expression dans la base temporelle

IV.3.3 - Expression dans le domaine fréquentiel

IV.3.4 - Expression dans le domaine des Z.

Chapitre 3 - SYNTHESE DES FILTRES NUMERIQUES RECURSIFS DANS LE DOMAINE TEMPOREL 35

I - APPROXIMATION D'UN FILTRE NUMERIQUE PAR UN FILTRE NUMERIQUE RECURSIF 35

I.1 - Filtre numérique récursif dont la réponse impulsionnelle possède K coordonnées non nulles 35

I.2 - Théorème fondamental des filtres numériques récursifs. 36

II - ETUDE DU PROBLEME DE LA DETERMINATION D'UN FILTRE NUMERIQUE RECURSIF DE REPONSE IMPULSIONNELLE DONNEE 36
II.1 - Position du problème 36
II.2 - Mise en équation du problème 37
II.3 - Similitude avec le problème de la déconvolution 38

III - NOTIONS MATHEMATIQUES RELATIVES AUX SYSTEMES LINEAIRES SURDETERMINES 40
III.1 - Généralités 40
III.2 - Solution approchée au sens de la minimisation de la norme Euclidienne. Pseudo-inverses 41
III.3 - Notion de conditionnement de la matrice d'un système linéaire 42

IV - APPLICATION A LA DETERMINATION DES COEFFICIENTS A PARTIR DE K COORDONNEES DE LA REPONSE IMPULSIONNELLE 43
IV.1 - Méthodes utilisées 43
IV.1.1 - Méthode de Burrus et Parks
IV.1.2 - Méthode directe
IV.2 - Résultats obtenus. 44

Chapitre 4 - PROBLEMES RELATIFS A LA REALISATION DES FILTRES NUMERIQUES RECURSIFS 47

INTRODUCTION

I - ANALYSE QUALITATIVE DES SOURCES D'ERREUR DANS UN FILTRE NUMERIQUE RECURSIF CABLE 47
I.1 - Erreur due à la quantification du signal d'entrée 47
I.2 - Erreur due à la quantification des paramètres 48
I.3 - Erreur d'arrondi dans les multiplications 48
I.4 - Erreurs et stabilité 48
I.4.1 - Erreurs n'affectant pas la stabilité
I.4.2 - Erreurs pouvant engendrer des instabilités
I.4.3 - Degré de stabilité d'un filtre numérique récursif.

II - DIFFERENTES STRUCTURES DE FILTRES NUMERIQUES RECURSIFS 49
II.1 - Structure directe 49
II.2 - Structure canonique 50
II.3 - Structure à éléments en cascade 51
II.4 - Structure à éléments en parallèle 51
II.5 - Remarque 52

III - ETUDE DE L'ERREUR DE QUANTIFICATION 52
III.1 - Généralités 52
III.2 - Quantification d'un signal certain 52
III.3 - Etude statistique de la quantification 53
III.3.1 - Principe de la méthode
III.3.2 - Quantification d'une variable aléatoire à une dimension
III.3.3 - Quantification d'une variable aléatoire à deux dimensions
III.3.4 - Application à la quantification d'un signal aléatoire
III.4 - Erreur due au nombre fini de niveaux 56
III.5 - Remarque 57

IV - CALCUL DE L'ERREUR A LA SORTIE D'UN FILTRE NUMERIQUE 57
IV.1 - Méthode de calcul 57
IV.2 - Expression de l'erreur en sortie 58
IV.3 - Application aux filtres numériques récursifs du 2ème ordre 59

V - ETUDE DE LA STABILITE D'UN FILTRE NUMERIQUE RECURSIF 60
V.1 - Généralités 60
V.2 - Stabilité en sensibilité au bruit de quantification 61
V.3 - Domaine de stabilité en fonction des coefficients 61
V.4 - Domaine de stabilité des filtres du 2ème ordre 62

VI - ASSOCIATION DES FILTRES EN PARALLELE OU EN CASCADE 64
VI.1 - Association en parallèle 64
VI.2 - Association en cascade 65

Chapitre 5 - PROJET DE REALISATION D'UN FILTRE NUMERIQUE RECURSIF PROGRAMMABLE 67

I - DESCRIPTION ET INTERET DE L'APPAREIL PROJETE 67

II - CHOIX DE LA STRUCTURE 67
II.1 - Choix du mode d'association 67
II.2 - Structure des éléments du 2ème ordre 68

III - ERREUR EN SORTIE ET STABILITE AVEC LA STRUCTURE CHOISIE 70
III.1 - Erreur en sortie 70
III.2 - Stabilité 71

IV - STRUCTURE TECHNOLOGIQUE DE L'APPAREIL 72
IV.1 - Algorithme de calcul 72
IV.2 - Schéma synoptique de l'appareil 73

Chapitre 6 - TEST D'UTILISATION D'UN FILTRE NUMERIQUE RECURSIF PROGRAMMABLE DANS UN PROBLEME DE FILTRAGE ADAPTE 77

I - CALCUL DES COEFFICIENTS DU FILTRE ADAPTE 77

II - TEST DU FILTRE OBTENU 78

CONCLUSION 79

Annexe I - Calcul de la fonction de Green d'un filtre numérique récursif 81

Annexe II - Algorithme de calcul d'un spectre. Etude de l'erreur d'estimation 83

Annexe III - Calcul des moments du 1er et du 2ème ordre d'une variable aléatoire quantifiée et du moment mixte de deux variables aléatoires quantifiées 85

Annexe IV - Approche d'un modèle de filtre donné dans le domaine temporel par des filtres numériques récursifs d'ordres croissants 89

BIBLIOGRAPHIE 95

NOTATIONS ET SYMBOLES UTILISES

$\mathcal{S}$ Espace des fonctions complexes d'une variable réelle indéfiniment dérivable et décroissant plus vite que $\frac{1}{t^\alpha}$ ainsi que toutes leurs dérivées.

$\mathcal{S}'$ Espace des distributions tempérées (dual topologique de $\mathcal{S}$).

$\mathcal{E}$ Espace des fonctions complexes d'une variable réelle, indéfiniment dérivables.

L^1 Espace des classes de fonctions complexes d'une variable réelle sommables.

L^2 Espace des classes de fonctions complexes d'une variable réelle de carré sommable [1].

BL^2 Espace des classes de fonctions complexes d'une variable réelle de carré sommable et à spectre borné [1].

$\int_\Delta$ Intégrale de Lebesgue sur le domaine $\Delta \subset R$

$*$ Opérateur de convolution.

sinct Fonction $\frac{\sin t}{t}$ prolongée par continuité au point $t = 0$

$\Pi_T(t)$ Fonction définie par : $\Pi_T(t) = 1 \quad t \in]-T,+T[$

$\Pi_T(t) = 0 \quad t \notin [-T,+T]$

non définie aux points $t = T$ et $t = -T$

$\mathcal{C}$ Opérateur de translation $\mathcal{C}_\tau[X(t)] = X(t-\tau)$

X^* Nombre complexe conjugué de X

$||X(t)||$ Norme de X(t) dans L^2

$\mathcal{E}'$ Espace des distributions à support borné (dual topologique de $\mathcal{E}$)

$\langle X,Y \rangle$ Produit scalaire de X par Y. Dans L^2 : $\langle X,Y \rangle = \int_R X(t)\, Y(t)\, dt$

$X(t) \rightleftharpoons x(\nu)$ X(t) a pour transformée de Fourier $x(\nu)$ et la transformation inverse existe.

δ_a Distribution de Dirac au point $t = a$

$\delta_{n,m}$ Symbole de Kronecker : $\delta_{n,m} = 1 \quad n = m$

$\delta_{n,m} = 0 \quad n \neq m$

---

[1] Nous appelons classe l'ensemble des fonctions qui ne diffèrent que sur un ensemble de mesure nulle.

ETUDE DU FILTRAGE NUMERIQUE RECURSIF DANS LE DOMAINE TEMPOREL
APPLICATION A LA SYNTHESE DES FILTRES ADAPTES

I N T R O D U C T I O N

Le filtrage numérique constitue un vaste domaine au sein des techniques de traitement numérique des signaux. Ces techniques prennent le pas sur les techniques analogiques en raison de leur précision et de leur souplesse. En effet, dans le traitement analogique on était directement asservi à un phénomène physique que l'on s'efforçait d'ajuster au traitement mathématique que l'on désirait faire subir au signal, ce qui n'est plus le cas pour le traitement numérique.

Le filtrage analogique a été abondamment étudié, mais le filtrage numérique, de conception récente, pose des problèmes assez nettement différents du fait des particularités inhérentes aux systèmes numériques, à savoir essentiellement l'échantillonnage et la quantification, qui nécessitent l'emploi de nouveaux outils mathématiques qui leurs sont adaptés (bases dénombrables dans les espaces fonctionnels de signaux, transformée en Z remplaçant la transformée de Laplace, etc...).

Dans les deux premiers chapitres nous nous sommes efforcés d'introduire de façon cohérente les notions théoriques de filtre numérique et de filtre numérique récursif en définissant au fur et à mesure les concepts mathématiques dont nous avions besoins.

Le troisième chapitre expose une méthode de calcul des coefficients d'un filtre numérique récursif à partir d'un gabarit donné dans le domaine temporel. On étudie différentes variantes de la méthode en comparant leurs performances.

Le quatrième chapitre étudie les problèmes relatifs à la réalisation technologique des filtres numériques récursifs, en particulier les questions de stabilité. Cette étude est utilisée au cinquième chapitre pour établir un projet concret de réalisation d'un filtre numérique récursif dont on peut programmer la réponse par affichage des coefficients.

Enfin, au chapitre 6, les performances de ce projet sont testées par simulation d'un problème de détection par filtrage adapté.

Chapitre 1

FORMALISME ET DEFINITION DES FILTRES NUMERIQUES

INTRODUCTION

Les filtres numériques, comme tous les appareils digitaux, sont des appareils effectuant des calculs sur des suites numériques.

Dans le cas du fonctionnement en temps réel qui est celui qui nous intéresse, la suite numérique d'entrée résulte en général de l'échantillonnage d'un signal continu que l'on désire traiter, tandis que la suite numérique de sortie constitue les échantillons du signal de sortie que l'on cherchera à reconstituer. C'est dans cette optique que nous nous placerons pour définir les filtres numériques [1].

Nous savons, d'après le théorème de Shannon, que seul les signaux à spectre borné sont échantillonnables sans perte d'information, c'est pourquoi nous commencerons par rappeler les principales propriétés de l'espace BL2 auquel appartiennent ces signaux.

I - RAPPEL DES PROPRIETES DE L'ESPACE BL2

I.1 - Définition de BL2

L'espace BL2 est l'ensemble des classes de fonctions complexes définies sur R, mesurables, de carré sommable sur R et à spectre uniformément borné [2].

---

1) Remarque : Les signaux que nous envisageons ici sont des signaux certains. Ceci est légitime puisque tout signal physique que nous désirons filtrer est connu au moment où il arrive sur l'entrée de l'appareil et peut donc être considéré comme un signal certain. Cependant, il arrive fréquemment que ce signal certain ne soit en fait qu'une réalisation particulière d'un processus aléatoire dont les propriétés de stationnarité et d'ergodicité nous autorisent à mener le traitement sur une seule réalisation pourvu qu'elle soit connue sur un intervalle de temps suffisamment long (infini en théorie). C'est là d'ailleurs tout l'intérêt de l'étude des processus aléatoires dans le traitement du signal. Toutefois, l'objet de la présente étude étant les filtres qui sont pour nous des éléments non aléatoires, le cadre de la théorie des signaux certains sera ici suffisant.

2) Ce qui signifie que la borne est indépendante de la fonction choisie.

Si X(t) est une fonction de la variable réelle t, $X(t) \in BL^2$ signifie :

1° - $\int_R |X(t)|^2 \, dt < \infty$, c'est-à-dire $X(t) \in L^2$, on a donc $BL^2 \subset L^2$

2° - La condition précédente permet d'affirmer que X(t) possède une transformée de Fourier x(ν) qui dans le cas où X(t) appartient également à L^1 est donnée par :

$$x(\nu) = \int_{-\infty}^{+\infty} X(t).e^{-2i\pi\nu t} \, dt$$

On sait que $x(\nu) \in L^2$ et que la transformation réciproque existe. La condition du spectre uniformément borné se traduira alors par :

$$\exists B > 0 \ , \quad \forall X(t) \ , \quad X(t) \in BL^2 \ , \quad |x(\nu)| = 0 \quad \text{si} \quad |\nu| > B$$

I.2 - Structure et propriétés de BL^2

I.2.1 - Propriétés des signaux de BL2

1° - On sait que le produit de la durée par la bande d'un signal est borné inférieurement. Il en résulte que les signaux de BL^2 dont le spectre est à support borné ont nécessairement une durée infinie : Tout signal X(t) de BL^2 a un support non borné.

2° - Le spectre de tout signal de BL^2 décroissant plus vite à l'infini que toute puissance de $\frac{1}{\nu}$, le signal qui est la transformée de Fourier inverse du spectre est indéfiniment dérivable, donc il est continu et borné, ainsi que ses dérivées en vertu du théorème de Bernstein. On peut donc énoncer :

Tout signal X(t) de BL^2 est borné, continu, indéfiniment dérivable et toutes ses dérivées sont bornées.

Plus précisément si le support du spectre de X(t) est [-B,+B] et M la borne supérieure de X(t) on a :

$$\left| \frac{d^n X(t)}{dt^n} \right| \leqslant M.(2\pi B)^n \qquad \text{(théorème de Bernstein).}$$

I.2.2 - BL2 est un espace de Hilbert

En effet :

1° - BL^2 est un sous espace vectoriel de L^2, puisque la transformation de Fourier étant linéaire la condition du spectre uniformément borné se conserve par transformation linéaire.

2° - BL^2 possède un produit scalaire et une norme qui sont ceux de L^2, à savoir : $\langle X,Y \rangle = \int_{-\infty}^{+\infty} X(t).Y^*(t) \, dt \quad ||X|| = (\langle X,X \rangle)^{1/2}$

3° - L^2 étant complet et $BL^2 \subset L^2$ toute suite de Cauchy dans BL^2 à une limite dans L^2. Or les fonctions de BL^2 étant à spectre uniformément borné, cette limite

est dans BL^2, en effet, nous avons par hypothèse :

$$\forall n,\ X_n(t) \in BL^2 \quad \text{et} \quad X_n(t) \to X(t), \qquad X(t) \in L^2$$

D'après le théorème de Parseval $||X_n(t)-X(t)|| = ||x_n(\nu)-x(\nu)||$

donc : $\forall \varepsilon\ \exists N,\ n > N \to ||x_n(\nu)-x(\nu)|| < \varepsilon$

or : $$||x(\nu)-x_n(\nu)||^2 = \int_{-B}^{+B} |x_n(\nu)-x(\nu)|^2\, d\nu + \int_{|\nu|>B} |x(\nu)|^2\, d\nu$$

par suite : $\forall \varepsilon$, $$\int_{|\nu|>B} |x(\nu)|^2\, d\nu < \varepsilon$$

d'où : $$\int_{|\nu|>B} |x(\nu)|^2\, d\nu = 0$$

soit enfin : $|x(\nu)| = 0 \qquad |\nu| > B$

donc : $x(\nu) \in BL^2$.

Il résulte de cette démonstration que BL^2 est complet. Il remplit donc les conditions pour avoir une structure d'espace de Hilbert.

I.2.3 - BL^2 est stable pour l'opération de convolution

Il s'agit de montrer que $X(t) \in BL^2$, $Y(t) \in BL^2 \to (X*Y) \in BL^2$.
En effet nous avons :

$$\left.\begin{array}{l} X(t)\in BL^2 \to x(\nu) \text{ est à support B. Borné} \\ Y(t)\in BL^2 \to y(\nu) \text{ est à support B. Borné} \end{array}\right\} \begin{array}{l} x(\nu).y(\nu) \text{ est à} \\ \text{support B. Borné} \end{array} \qquad (1)$$

$$(1) \to x(\nu).y(\nu) \in L^2 \to X*Y \in L^2 \qquad (2)$$

$$(1) \text{ et } (2) \to X*Y \in BL^2$$

I.2.4 - La convergence dans BL^2

Comme dans tout espace de Hilbert il existe dans BL^2 deux sortes de convergences.

- La convergence faible qui se définit par :

$$\forall g \in BL^2\ ,\ \forall \varepsilon\ ,\ \exists N\ ,\ \forall n\ ,\ n > N \to |<f_n,g>-<f,g>| < \varepsilon$$

- la convergence forte qui dans BL^2 n'est autre que la convergence en moyenne quadratique et qui se définit par :

$$\forall \varepsilon\ ,\ \exists N\ ,\ \forall n\ ,\ n > N \to ||f_n-f|| < \varepsilon$$

Dans tout ce qui va suivre, c'est ce dernier type de convergence que nous

utiliserons, sans préciser convergence '"forte".

I.3 - La base temporelle de BL^2

Nous allons montrer que l'ensemble des signaux :

$S_n(t) = \sqrt{2B}.\text{sinc}\left|2\pi B(t-\frac{n}{2B})\right|$ pour $n\in Z$ constitue une base de Hilbert de BL^2, c'est-à-dire une base orthonormée totale dans BL^2.

1° - Les $S_n(t)$ forment un système orthonormé.

En effet, soit $s_n(\nu)$ la transformée de Fourier de $S_n(t)$ nous avons, en appliquant le théorème de Parseval :

$$<S_n(t).S_m(t)> = <s_n(\nu).s_m(\nu)>$$

$$<S_n(t).S_m(t)> = <\frac{1}{\sqrt{2B}}\Pi_B(\nu).e^{-2i\pi\nu\frac{n}{2B}} , \frac{1}{\sqrt{2B}}\Pi_B(\nu)\ e^{-2i\pi\nu\frac{m}{2B}}>$$

$$<S_n(t).S_m(t)> = \frac{1}{2B}\int_{-B}^{+B} e^{-2i\pi\frac{n-m}{2B}\nu}\,d\nu$$

$$<S_n(t).S_m(t)> = \delta_{n,m}$$

2° - Les $S_n(t)$ forment un système total dans BL^2, c'est-à-dire que tout élément de BL^2 est la limite d'une suite de combinaisons linéaires finies de fonctions $S_n(t)$ [3].

En effet : $X(t)\in BL^2 \rightarrow x(\nu)$ est à spectre B. Borné

donc on peut écrire : $x(\nu) = x_p(\nu).\Pi_B(\nu)$ (3)

où $x_p(\nu)$ est une fonction périodique de période 2B qui est presque partout égale à $x(\nu)$ sur $[-B,+B]$. Il en résulte que $x_p(\nu)\ \in L^2[-B,+B]$ et peut donc se décomposer sur la base de Fourier de cet espace, c'est-à-dire que l'on a :

$$x_p(\nu) = \sum_{n=-\infty}^{+\infty} a_n.e^{-2\pi i\frac{n}{2B}\nu} \qquad (4)$$

avec :

$$a_n = 2B\int_{-B}^{+B} x_p(\nu).e^{2\pi i\frac{n}{2B}\nu}$$

$$a_n = 2B\int_{-\infty}^{+\infty} x(\nu).e^{2\pi i\frac{n}{2B}\nu}$$

$$a_n = <x(\nu),\ e^{-2\pi i\frac{n}{2B}\nu}$$

---

3) La démonstration qui suit est la démonstration classique du théorème d'échantillonnage. Nous la reprendrons néanmoins en la remplaçant dans le cadre de l'exposé.

Cette dernière expression est celle d'un produit scalaire entre $x(\nu) \in \mathcal{E}'$ (espace des distributions à support borné)
et $e^{-2\pi i \frac{n}{2B}\nu} \in \mathcal{E}$ (espace des fonctions indéfiniment dérivables).

Par transformation de Fourier, $x(\nu)$ donne $X(t) \in \mathcal{E}$ et $e^{-2\pi i \frac{n}{2B}\nu}$ donne $\delta_{\frac{n}{2B}} \in \mathcal{E}'$.

D'après le théorème de Parseval, on a donc :

$$a_n = 2B. \langle X(t). \delta_{\frac{n}{2B}} \rangle = 2B.X\left(\frac{n}{2B}\right)$$

d'où, compte tenu de (3) et (4) :

$$x(\nu) = \sum_{n=-\infty}^{n=+\infty} 2B.X\left(\frac{n}{2B}\right) .\Pi_B(\nu).e^{-2\pi i \frac{n}{2B}\nu}$$

c'est-à-dire par transformée de Fourier :

$$X(t) = \sum_{n=-\infty}^{n=+\infty} X\left(\frac{n}{2B}\right).\text{sinc } 2\pi B \left(t-\frac{n}{2B}\right)$$

ou : $$X(t) = \sum_{n=-\infty}^{n=+\infty} \frac{1}{\sqrt{2B}} .X\left(\frac{n}{2B}\right).S_n(t)$$

Tout élément $X(t)$ de BL^2 est donc la limite quand $N \to \infty$ de la suite

$$X_N(t) = \sum_{n=-N}^{n=+N} \frac{1}{\sqrt{2B}} . X\left(\frac{n}{2B}\right) . S_n(t)$$

Les $S_n(t)$ constituent donc une base de Hilbert de l'espace BL^2 et l'intérêt de cette base est que les coordonnées d'un signal quelconque de BL^2 ne sont autres que les valeurs du signal échantillonné à la cadence $T = \frac{1}{2B}$ (au coefficient $\frac{1}{\sqrt{2B}}$ près).

II - DEFINITION DES FILTRES NUMERIQUES

II.1 - Les transformations linéaires dans BL^2

II.1.1 - Définition

Nous appellerons transformation linéaire dans L^2 (donc aussi dans BL^2) toute transformation T qui possède les trois propriétés suivantes :

1° T est additive : $\forall f, \forall g \in BL^2 \quad T(f+g) = Tf + Tg$

2° T est homogène : $\forall f \in BL^2, \forall \lambda \in C \quad T(\lambda f) = \lambda.T(f)$

3° T est continue : $f_n \to f \to Tf_n \to Tf$
au sens de la convergence dans BL^2.

II.1.2 - Transformations bornées

Une transformation additive et homogène est dite bornée s'il existe un nombre réel positif M tel que : $\forall f \in BL^2 \quad ||Tf|| \leq M.||f||$.

On démontre [16] qu'une transformation additive, homogène est bornée si et seulement si elle est continue. Il en résulte que les transformations linéaires, au sens défini précédemment sont des transformations bornées.

II.1.3 - Convergence des transformations linéaires dans BL^2

Il existe plusieurs sortes de convergence des transformations linéaires, nous utiliserons celle qui se rapporte à la convergence en moyenne des signaux de BL^2 laquelle possède le double avantage d'avoir un sens physique clair et d'être facilement manipulable dans les calculs.

Nous dirons donc qu'une suite $T_n (n \in N)$ de transformations linéaires converge vers T si pour tout signal f(t) de BL^2, $T_n f$ converge vers Tf en moyenne quadratique (c'est-à-dire $||T_n f - Tf|| \to 0$)

II.2 - Les filtres linéaires dans BL^2

II.2.1 - Les translations

On appellera translation d'amplitude τ la transformation $\mathcal{C}_\tau$ définie par :

$$\forall X(t) \in BL^2 \qquad \mathcal{C}_\tau X = X(t-\tau)$$

On montre aisément que les translations sont des transformations linéaires de BL^2 dans BL^2, au sens défini dans le paragraphe II.1.1.

II.2.2 - Les filtres linéaires [4]

Nous appellerons filtre linéaire toute transformation linéaire qui commute avec les translations.

$\mathcal{F}$ est un filtre linéaire dans BL^2 si les deux conditions suivantes sont remplies :

1° - $\mathcal{F}$ est une transformation linéaire de BL^2 dans BL^2

2° - $\forall X(t) \in BL^2, \quad \mathcal{F}\mathcal{C}_\tau X = \mathcal{C}_\tau \mathcal{F} X$

c'est-à-dire si $Y = \mathcal{F} X$, $Y(t-\tau) = \mathcal{F} X(t-\tau)$.

II.2.3 - Théorème fondamental des filtres linéaires

L'expression courante, des filtres linéaires se fonde sur le théorème suivant dû à Laurent Schwartz :

---

[4] Ce que nous appelons ici simplement filtre linéaire se rencontre souvent dans la littérature le nom de filtre linéaire stationnaire.

A toute transformation linéaire $X(t) \multimap Y(t') = \mathcal{F} X(t)$ qui commute avec les translations, on peut faire correspondre une distribution F, telle que l'on ait : $Y = F*X$

Dans le cas où $X = \delta$ (distribution de Dirac), on a : $Y = F*\delta = F$. C'est pourquoi la distribution F est appelée réponse impulsionnelle du filtre $\mathcal{F}$.

Dans le cas qui nous intéresse X et Y sont des éléments de BL^2. Nous allons maintenant étudier un sous ensemble de l'ensemble des Filtres linéaires de BL^2 qui sera défini par le fait que la réponse impulsionnelle F appartiendra à un sous espace de l'espace des distributions, à savoir BL^2 lui-même.

II.3 - Les filtres numériques dans BL^2

II.3.1 - Définition

Le produit de convolution de deux éléments quelconques de BL^2 étant encore un élément de BL^2 (§. I.2.2), on peut poser la définition suivante :

Nous appellerons filtre numérique toute transformation qui à un signal X(t) de BL^2 fait correspondre le signal Y(t) de BL^2 donné par : $Y(t) = H*X$ où H(t) est un élément de BL^2 appelé réponse impulsionnelle du filtre numérique.

Il résulte de cette définition que l'ensemble des filtres numériques est isomorphe à l'ensemble des éléments de BL^2 et par suite, nous pourrons les identifier. Ainsi, tout élément de BL^2 pourra représenter soit un signal, soit un filtre dont il est la réponse impulsionnelle.

II.3.2 - Convergence des filtres numériques

Il résulte de ce qui précède que nous pourrons définir la norme d'un filtre numérique comme la norme dans BL^2 de sa réponse impulsionnelle. Nous définirons ensuite la convergence d'une suite de filtres par la convergence en moyenne quadratique de la suite des réponses impulsionnelles. Nous écrirons donc :

$$\lim_{n\to\infty} \mathcal{H}_n = \mathcal{H} \leftrightarrow \lim_{n\to\infty} ||H_n - H|| = 0$$

Nous allons montrer que cette notion de convergence entraîne la convergence en moyenne quadratique des signaux de sortie, au sens défini au paragraphe II.1.3.

Soit $H(t) \in BL^2$ la réponse impulsionnelle d'un filtre numérique et $X(t) \in BL^2$ un signal quelconque ; nous avons :

$$||H*X|| = ||h(\nu).x(\nu)|| = \int_{-B}^{+B} h^2(\nu).x^2(\nu)\, d\nu \qquad (5)$$

Nous nous bornerons dans la démonstration au cas courant où X(t) est sommable, dans ces conditions on sait que $x(\nu)$ est borné, donc également $x^2(\nu)$. Soit M_{x2} la borne supérieure de $x^2(\nu)$ on aura d'après (5) :

$$||H*X|| \leqslant 2B.M_{x2}.||H||$$

Par suite, $||H_n*X - H*X|| = ||(H_n - H)*X|| \leqslant 2B.M_{x2}.||H_n - H||$

donc $\lim_{n\to\infty} ||H_n - H|| = 0 \rightarrow \lim_{n\to\infty} ||H_n * X - H * X|| = 0.$

Ce qui se traduit physiquement en disant que si deux filtres numériques H et H' sont voisins au sens où $||H-H'||$ est petit, les signaux de sortie Y et Y' correspondant au même signal d'entrée X sont voisins en moyenne quadratique, c'est-à-dire que $||Y-Y'||$ est petit.

C'est cette propriété qui nous autorisera ultérieurement à utiliser la valeur de $||H-H'||$ comme évaluation de la "proximité" de deux filtres.

II.3.3 - Expression du filtrage numérique dans la base temporelle de BL^2

Nous allons exprimer la relation fondamentale du filtrage : $Y = H*X$ dans la base temporelle de BL^2, ce qui est possible, puisque Y, H et X sont des éléments de BL^2.

Nous avons donc sur cette base les décompositions suivantes :

$$X(t) = \sum_{n=-\infty}^{+\infty} X_n . S_n(t) \qquad X_n = \frac{1}{\sqrt{2B}} . X\left(\frac{n}{2B}\right)$$

$$Y(t) = \sum_{n=-\infty}^{+\infty} Y_n . S_n(t) \qquad Y_n = \frac{1}{\sqrt{2B}} . Y\left(\frac{n}{2B}\right)$$

Nous savons que $S_p * S_q$ a pour transformée de Fourier le produit $s_p(\nu).s_q(\nu)$.

Or : $s_p(\nu) = \frac{1}{\sqrt{2B}} \Pi_B(\nu).e^{-2i\pi\frac{p}{2B}\nu}$

donc : $s_p(\nu).s_q(\nu) = \frac{1}{2B} \Pi_B(\nu).e^{-2i\pi\frac{p+q}{2B}\nu}$

d'où : $S_p * S_q = \frac{1}{\sqrt{2B}} S_{p+q}(t)$

Compte tenu de cette relation et portant les décompositions de Y(t), X(t) H(t) sur la base temporelle dans l'expression $Y = H*X$ il vient en identifiant la relation suivante entre les coordonnées de Y, H et X :

$$Y_n = \frac{1}{\sqrt{2B}} \sum_{p+q=n} H_q . X_p$$

Cette expression peut encore s'écrire :

$$Y_n = \frac{1}{\sqrt{2B}} \sum_{m=-\infty}^{+\infty} H_{n-m} . X_m \qquad (6)$$

ou : $$Y_n = \frac{1}{\sqrt{2B}} \sum_{m=-\infty}^{+\infty} H_m . X_{n-m} \qquad (6')$$

Ces deux relations équivalentes définissent l'expression du filtrage numérique dans la base temporelle.

On voit que la valeur des échantillons $Y(\frac{n}{2B})$ du signal de sortie se calcule directement à partir des échantillons $X(\frac{n}{2B})$ du signal d'entrée, d'où le nom de filtre numérique donné à cette classe de filtres.

II.3.4 - Exemples de filtres numériques

II.3.4.1 - Filtre numérique identité

C'est le filtre qui a pour composantes : $H_n = \frac{1}{\sqrt{2B}} \delta_{n,o}$ dans la base temporelle. En effet, en portant cette valeur de la suite H_n dans (6), il vient pour tout n : $Y_n = X_n$ d'où $Y(t) = X(t)$.

La réponse impulsionnelle du filtre numérique identité est donc :

$$\Delta(t) = \text{sinc } 2\pi Bt$$

Remarquons que pour les signaux de L^2 le filtre identité a pour réponse impulsionnelle non pas une "fonction" de L^2 mais la distribution de Dirac δ. Ainsi, l'équivalent d'une impulsion dans l'espace BL^2 est le signal $X(t) = \text{sinc } 2\pi Bt$.

II.3.4.2 - Filtres numériques de translation

Considérons les filtres $\mathcal{C}_p$ qui ont pour composants dans la base temporelle : $H_n = \frac{1}{\sqrt{2B}} \delta_{n,p}$ $(n,p \in Z)$. Portant dans (6'); il vient $Y_n = X_{n-p}$ pour tout n, c'est-à-dire :

$$Y(t) = X(t-\frac{p}{2B})$$

Ainsi, le résultat du filtrage par un tel filtre est une translation de la valeur $\frac{p}{2B}$ $(p \in Z)$. Ce filtre a pour réponse impulsionnelle :

$$T_p(t) = \text{sinc } 2\pi B \left(t - \frac{p}{2B}\right)$$

Remarquons que les filtres de translation d'une valeur τ qui n'est pas de la forme $\frac{p}{2B}$ $(p \in Z)$ ont pour réponse impulsionnelle la distribution de Dirac au point $t = \tau$ soit δ_τ. Ce ne sont donc pas des filtres numériques.

III.- FILTRES NUMERIQUES PHYSIQUEMENT REALISABLES

III.1 - Les signaux physiques

Les signaux physiques que nous observons ne peuvent être que des approximations des représentations mathématiques que nous utilisons. En effet, le temps d'observation d'un phénomène étant nécessairement fini, les signaux que nous manipulons sont en fait des signaux à support borné, dont le spectre s'étend nécessairement jusqu'à l'infini ; or tous les appareils physiques que nous utilisons pour observer ces signaux, se comportent comme des filtres passe bas et par suite les déforment en coupant les fréquences hautes de leur spectre.

III.2 - Les signaux dans les systèmes digitaux

Les systèmes digitaux ne sont capables de traiter que des suites de nombres. Lorsqu'on utilise ces appareils pour traiter des signaux, on doit associer à chaque signal une suite numérique qui constitue les coordonnées du signal dans une base dénombrable. Le cas le plus fréquent est celui où l'on se limite au traitement des signaux de BL^2 que l'on représente par leurs coordonnées dans la base temporelle de BL^2 ainsi que nous l'avons exposé précédemment.

En fait, les suites numériques traitées sont finies, c'est-à-dire que le nombre des éléments, non nuls de ces suites doit être fini.

Cependant, considérant que le traitement numérique d'un signal peut se prolonger aussi longtemps qu'on le désire, nous serons moins restrictifs en définissant les signaux réellement traitables par un système digital comme les signaux de BL^2 dont les coordonnées dans la base temporelle sont toutes nulles en deça d'un certain rang n que l'on peut toujours choisir comme origine (n=0).

On peut montrer très aisément que les signaux définis forment un sous espace vectoriel de BL^2 que nous noterons BL^2_+.

En résumé, les signaux traitables par des filtres numériques réels sont les signaux de l'espace BL^2_+ défini de la manière suivante :

$$X(t)\in BL^2_+ \leftrightarrow \begin{cases} X(t)\in BL^2 \\ \forall n,\ n\in Z,\ n<0,\ X\left(\frac{n}{2B}\right)=0 \end{cases}$$

Remarquons que les signaux de BL^2_+ ne sont en général pas nuls pour les valeurs négatives de t qui ne sont pas de la forme $\frac{n}{2B}$.

III.3 - La causalité pour les filtres numériques

Les filtres numériques que nous envisageons ici étant destinés à travailler en temps réel, seuls sont réalisables ceux qui satisfont au critère de causalité à savoir que l'échantillon d'ordre n du signal de sortie ne peut être fonction que des échantillons d'ordre inférieur ou égal à n du signal d'entrée.

L'expression (6) nous permet d'exprimer mathématiquement cette condition sous la forme :

$$\forall X(t),\ \sum_{m=-\infty}^{+\infty} H_{n-m}\ X_m \equiv \sum_{m=-\infty}^{n} H_{n-m}\ X_m$$

ce qui entraîne $\forall m,\ m > n \quad H_{n-m} = 0$.

C'est-à-dire que H(t) doit appartenir à l'espace BL^2_+ :

Les filtres numériques répondant au critère de causalité sont des filtres numériques dont la réponse impulsionnelle appartient à l'espace BL^2_+.

Remarque : Cette notion de causalité ne se confond pas exactement avec celle correspondant aux signaux continus : en effet, $H(t)\in BL^2_+$ n'entraîne pas la nullité de H(t) pour toute valeur négative de t ($t\in R$).

Il est vrai que le signal impulsion dans BL^2 et dont H(t) est la réponse en sortie du filtre est lui-même non nul pour tout t négatif puisqu'il s'agit de la fonction $\Delta(t) = \text{sinc } 2\pi Bt$.

III.4 - L'espace BL_+^2 des filtres numériques et des signaux physiques réels

Nous avons vu que les signaux effectivement traitables par des appareils digitaux tels que les filtres numériques sont des signaux de l'espace BL_+^2. De même nous venons de voir que les filtres numériques répondant au critère de causalité ont leur réponse impulsionnelle dans l'espace BL_+^2.

Ainsi, de la même manière que BL^2 était l'espace des signaux et des filtres numériques au sens général, son sous espace vectoriel BL_+^2 est l'espace des signaux physiques échantillonnables et des filtres numériques réalisables en temps réel ; tout élément de BL_+^2 désignant soit un signal soit une réponse impulsionnelle attachée à un filtre.

- le filtre identité, de réponse impulsionnelle :

$$\Delta(t) = \text{sinc } 2\pi Bt$$

appartient à l'espace BL_+^2 ;

- les filtres de translation, de réponse impulsionnelle :

$$T_p(t) = \text{sinc } 2\pi B\left(t-\frac{p}{2B}\right)$$

appartiennent à BL_+^2 si p est positif.

C'est la traduction du fait que seul le retard (et non l'avance) est réalisable en temps réel.

Enfin, nous noterons que pour des signaux et des filtres de BL_+^2 les expressions (6) et (6') donnant les échantillons du signal de sortie en fonction de ceux du signal d'entrée d'un filtre numérique s'écrivent respectivement :

$$Y_n = \frac{1}{\sqrt{2B}} \sum_{m=o}^{m=n} H_{n-m} \cdot X_m \qquad (7)$$

$$Y_n = \frac{1}{\sqrt{2B}} \sum_{m=o}^{m=n} H_m \cdot X_{n-m} \qquad (7')$$

Chapitre 2

LES FILTRES NUMERIQUES RECURSIFS

I - EQUATION AUX DIFFERENCES DEFINISSANT UN FILTRE NUMERIQUE RECURSIF

I.1 - Cas des systèmes continus

Nous savons qu'il existe des systèmes linéaires dont le comportement est décrit par une équation différentielle linéaire à coefficients constants. Il en est ainsi par exemple pour les systèmes mécaniques et électriques dits à "constantes localisées" et invariants dans le temps. L'équation différentielle décrivant ces systèmes relie une grandeur d'entrée qui est en général une fonction du temps appelée excitation, à une grandeur de sortie appelée "réponse" du système.

Dans le cas des systèmes initialement au repos, c'est-à-dire avec des conditions initiales nulles, à toute excitation correspond une réponse unique cette correspondance étant linéaire et invariante par translation dans le temps, nous avons donc affaire à un filtre linéaire particulier définit par une équation différentielle linéaire à coefficients constants avec des conditions initiales nulles.

Nous allons montrer qu'il en va de même pour les systèmes échantillonnés : on peut définir une classe particulière de filtres numériques à l'aide d'équations aux différences finies linéaires et à coefficients constants, ce sont ces filtres que nous appellerons filtres numériques récursifs.

I.2 - Suite numérique solution d'une équation aux différences finies

Soit $\{X_k\}$ $(k \in Z)$ une suite connue de nombres réels ou complexes
et $\{Y_k\}$ $(k \in Z)$ une suite inconnue.

Soit l'équation aux différences finies suivantes :

$$k \in Z \qquad \sum_{n=0}^{N} b_n . Y_{k-n} = \sum_{m=0}^{M} a_m . X_{k-m} \qquad (8)$$

I.2.1 - L'application $\{X_k\} \rightarrow \{Y_k\}$

Nous savons que si l'on se donne arbitrairement N nombres λ_i appelés conditions initiales, il existe une suite $\{Y_k\}$ et une seule qui soit solution de l'équation (8) et telle que : $Y_k = \lambda_k$ pour $k = q, q+1, \ldots, q+N-1$.

Ainsi pour λ_i et q fixés, à toute suite $\{X_k\}$ correspond une suite unique $\{Y_k\}$ solution de (8) avec les conditions initiales fixées. Nous avons ainsi défini une application dans l'ensemble des suites numériques.

Nous envisagerons désormais le cas de l'application $\mathcal{H}$ définie avec q = -N, λ_i = o c'est-à-dire les conditions initiales : Y_k = o pour k = -1, -2,...,-N que nous appellerons conditions initiales homogènes.

L'application $\mathcal{H}$ définie ci-dessus est alors linéaire.

En effet, l'équation (8) étant linéaire la suite $\{Y_k\}=\alpha_1 \{Y_{1k}\} + \alpha_2\{Y_{2k}\}$ est la solution correspondant à la suite $\{X_k\} = \alpha_1\{X_{1k}\} + \alpha_2\{X_{2k}\}$ sous réserve que la relation $\lambda_i = \alpha_1 \lambda_{1i} + \alpha_2 \lambda_{2i}$ (i=q,q+1,...,q+N-i) entre les conditions initiales respectives des suites $\{Y_k\}$, $\{Y_{1k}\}$ et $\{Y_{2k}\}$ soit vérifiée. Or dans le cas des conditions initiales homogènes, cette relation est vérifiée quels que soient α_1 et α_2 puisque $\lambda_i = \lambda_{1i} = \lambda_{2i} = 0$ pour i = -1, -2,..., -N.

I.2.2 - Cas des suites à termes nuls pour k ≤ 0 (ensemble S_+)

Restreignons l'ensemble de départ de l'application précédente l'ensemble S_+ des suites $\{X_k\}$ telles que $\forall k, k\in Z, k<0 \rightarrow X_k=0$.

L'équation (8) peut se mettre sous la forme :

$$Y_{k-N} = -\frac{1}{b_N}\sum_{n=o}^{N-1} b_n.Y_{k-n} + \sum_{m=o}^{M} a_m.X_{k-m} \qquad (8')$$

en utilisant cette relation, on montre aisément par récurrence que les conditions initiales homogènes entraînent que l'image de toute suite de S_+ par l'application $\mathcal{H}$ est encore une suite de S_+.

I.2.3 - Commutativité de $\mathcal{H}$ avec les translations dans S_+

Nous appellerons translation $\mathcal{C}_p$ dans S_+ l'application linéaire qui à la suite $\{U_k\}$ de S_+ fait correspondre la suite $\{V_k\} = \{U_{k-p}\}$ de S_+ avec p > o (p∈Z).

Considérons l'équation (8) avec $X_k \in S_+$ cette équation devant être vérifiée pour tout k∈Z on peut y remplacer k par k-p avec p > o. D'autre part, nous avons vu qu'avec les conditions initiales homogènes, $\{X_k\}\in S_+ \rightarrow \{Y_k\}\in S_+$, donc Y_{k-p} = o pour k=-1, -2,..., -N puisque p est positif.

En définitive, on a donc :

$$\sum_{n=o}^{N} b_n.Y_{k-p-n} = \sum_{m=o}^{M} a_m.X_{k-p-m}$$

$$Y_{k-p} = 0 \qquad k = -1, -2,..., -N$$

Ceci nous montre que la suite $\mathcal{C}_p\{Y_k\} = \{Y_{k-p}\}$ est l'image par l'application $\mathcal{H}$ de la suite $\mathcal{C}_p\{X_k\} = \{X_{k-p}\}$. Nous avons donc démontré la propriété suivante :

$\{X_k\} \in S_+$, $\{Y_k\} \in S_+$

$p \in Z$, $p > o$

$$\mathcal{C}_p \mathcal{H}\{X_k\} = \mathcal{H}\mathcal{C}_p\{X_k\}$$

L'opérateur $\mathcal{H}$ commute avec les translations dans S_+.

I.2.4 - Expression directe de l'application $\mathcal{H}$

Désignons par $\{\Delta_k\}$ la suite dont les termes Δ_k sont définis par : $k\in Z$, $\Delta_k = \delta_{k,o}$ et soit $\{H_k\}$ la suite image de $\{\Delta_k\}$ par l'application $\mathcal{H}$.

Nous avons donc $\{H_k\} = \mathcal{H}\{\Delta_k\}$ et $\{\Delta_k\}\in S_+ \rightarrow \{H_k\}\in S_+$. Pour tout $k\in Z$ on a la relation évidente suivante : $X_k = \sum_{p=o}^{k} \delta_{k-p}.X_p$ d'où l'on déduit, compte tenu de la définition de la suite $\{\Delta_k\}$:

$$\{X_k\} = \sum_{p=o}^{k} X_p.\mathcal{C}_p\{\Delta_k\} \qquad (9)$$

$\mathcal{C}_p$ désignant la translation d'ordre p dans S_+, définie au paragraphe précédent. Appliquant aux deux membres de l'expression (9) l'opérateur $\mathcal{H}$, il vient, compte tenu de la linéarité de $\mathcal{H}$:

$$\mathcal{H}\{X_k\} = \sum_{p=o}^{k} X_p.\mathcal{H}\mathcal{C}_p\{\Delta_k\}$$

c'est-à-dire en utilisant la commutativité de $\mathcal{H}$ et $\mathcal{C}_p$:

$$\mathcal{H}\{X_k\} = \sum_{p=o}^{k} X_p.\mathcal{C}_p\,\mathcal{H}\{\Delta_k\}$$

ou encore : $$\{Y_k\} = \sum_{p=o}^{k} X_p.\mathcal{C}_p\{H_k\}$$

c'est-à-dire que les termes Y_k de la suite $\{Y_k\}$ image de $\{X_k\}$ par l'application $\mathcal{H}$ s'expriment par la relation fondamentale suivante :

$$Y_k = \sum_{p=o}^{k} H_{k-p}.X_p \qquad (10)$$

qui peut aussi se mettre sous la forme équivalente :

$$Y_k = \sum_{p=o}^{k} H_p.X_{k-p} \qquad (10')$$

I.2.5 - Caractérisation de l'application $\mathcal{H}$

L'équation (8) étant définie à un facteur multiplicatif près nous pouvons par exemple poser $b_o = 1$. Il en résulte que l'application $\mathcal{H}$ est caractérisée par la donnée des M+N+1 paramètres : $b_n (n=1,2,\ldots,N)$ et $a_m (m=0,1,\ldots,M)$.

D'après l'expression (10), la donnée de la suite $\{H_k\}$ image de $\{\Delta_k\}$ par $\mathcal{H}$ suffit à déterminer l'application .

Cependant, nous verrons au chapitre suivant que si à la donné de tout ensemble de coefficients (b_n, a_m) correspond une suite $\{H_k\}$ qui détermine aussi bien $\mathcal{H}$, la réciproque n'est pas vraie : à toute suite $\{H_k\}\in S_+$ définissant une application $\mathcal{H}$ dans S_+ par l'expression (10) ne correspond pas nécessairement un ensemble fini de coefficients (b_n, a_m) pour l'équation (8).

I.3 - Définition des filtres numériques récursifs

D'après les résultats du premier chapitre, toute suite $\{X_k\}\in S_+$ et telle que la série $\sum_{k=o}^{\infty} |X_k|^2$ soit convergente peut être considérée comme la suite des coordonnées dans la base temporelle d'un signal X(t) de l'espace BL_+^2. Il y a donc correspondance biunivoque entre les éléments de BL_+^2 et ceux de l'espace SL_+^2 formé par les suites de S_+ dont la série des modules carré des termes converge.

Considérons les applications $\mathcal{H}$ définies au chapitre I.2.4 par la suite $\{H_k\}$ et restreignons nous au cas des suites $\{H_k\}\in SL_+^2$. D'après l'expression (10), nous pouvons affirmer que toute suite $\{H_k\}\in SL_+^2$ définit une application linéaire de SL_+^2 dans SL_+^2 permutable avec les translations. Il en résulte, en vertu de l'isomorphisme entre BL_+^2 et SL_+^2 que l'on définit ainsi une application linéaire permutable avec les translations de BL_+^2 dans BL_+^2. Comparant les expressions (10) et (7) (dans lesquelles on a posé 2B = 1 ce qui revient à prendre par unité de temps le pas d'échantillonnage) on en déduit que la suite $\{H_k\}\in SL_+^2$ définit un filtre numérique causal, les H_k n'étant autres que les coordonnées dans la base temporelle de la réponse impulsionnelle $H(t)\in BL_+^2$ de ce filtre.

On a donc l'énoncé suivant :

Théorème et définition :

Toute équation aux différences finies du type (8) avec des conditions initiales homogènes et telles que pour $X_k = \delta_{k,o}$, $\{Y_k\} = \{H_k\}$ soit une suite pour laquelle la série de terme général $|H_k|^2$ converge, définit un filtre numérique causal dont les coordonnées de la réponse impulsionnelle dans la base temporelle de BL_+^2 sont les H_k. Un tel filtre numérique sera appelé filtre numérique récursif.

II - INTERET TECHNOLOGIQUE DES FILTRES NUMERIQUES RECURSIFS

Ainsi que nous l'avons vu précédemment tout filtre numérique causal est régit par une équation du type Y = H*X avec Y, H, $X\in BL_+^2$, ce qui dans la base temporelle s'écrit :

$$Y_n = \frac{1}{\sqrt{2B}} \sum_{m=o}^{m=n} H_{n-m} \cdot X_m \qquad (7)$$

On peut envisager un calculateur spécialisé (convolueur) qui calculerait les échantillons du signal de sortie à partir de ceux du signal d'entrée en utilisant l'expression (7). Le calculateur présente l'inconvénient pour calculer l'échantillon du signal de sortie correspondant à un instant donné, de nécessiter l'accumulation de tous les échantillons du signal d'entrée obtenus jusqu'à cet instant.

Nous pouvons envisager également un calculateur qui fournirait les échantillons du signal de sortie à partir de ceux du signal d'entrée en utilisant l'équation (8) des filtres numériques récursifs que l'on peut écrire sous la forme suivante :

$$Y_n = \sum_{m=o}^{M} a_m . X_{k-m} - \frac{1}{b_o} \sum_{n=1}^{N} b_n . Y_{k-n}$$

Nous constatons cette fois que les sommations ne portent que sur un nombre fini et fixe de termes, d'où la possibilité d'une réalisation technologique simple. Le stockage d'un certain nombre d'échantillons de sortie nous permet ainsi de nous affranchir de la connaissance de tous les échantillons du signal d'entrée depuis son origine (d'où le nom de "récursif" donné à ce type de filtre numérique).

Il n'en reste pas moins que les filtres numériques récursifs ne constituent qu'une classe particulière de filtres numériques. Le problème se pose donc, étant donné un filtre numérique donné quelconque, de savoir s'il est possible de l'approcher convenablement par un filtre numérique récursif plus aisément réalisable technologiquement. Cette étude fera l'objet du chapitre suivant.

III - DETERMINATION DE LA REPONSE IMPULSIONNELLE A PARTIR DES COEFFICIENTS CONDITION DE STABILITE

III.1 - Equation de convolution définissant un filtre numérique récursif

Soient $\mathcal{C}_X$ et $\mathcal{C}_Y$ les filtres numériques dont les réponses impulsionnelles sont les suivantes : $T_X(t) = \sum_{m=o}^{M} a_m . T_m(t)$ et $T_Y(t) = \sum_{n=o}^{N} b_n . T_n(t)$ où les $T_m(t)$ et $T_n(t)$ sont des réponses impulsionnelles de filtres de translation, de la forme : $T_p(t) = \text{sinc } 2\pi B \left(t - \frac{p}{2B}\right)$. $X(t)$ étant un signal de BL^2_+, considérons l'équation de convolution suivante où $Y(t)$ est un signal inconnu de BL^2_+ :

$$T_Y * Y = T_X * X \qquad (11)$$

Cette équation en projection dans la base temporelle de BL^2 n'est autre que l'équation aux différences (8) avec des conditions initiales homogènes puisque nous restreignons à BL^2_+ l'espace des solutions de (11).

Ainsi, sous réserve qu'elle possède une solution dans BL^2_+, l'équation (11) est équivalente à l'équation (8) avec des conditions initiales homogènes, elle définit donc un filtre numérique récursif.[5)]

---

5) Nous savons que cette solution est unique, si elle existe dans BL^2 puisqu'elle équivaut à la solution d'une équation aux différences finies d'ordre N avec N conditions initiales nulles.

La réponse impulsionnelle $H(t) \in BL^2$ de ce filtre s'obtient en prenant pour signal d'entrée $X(t) = \Delta(t)$ avec $\Delta(t) = \text{sinc } 2\pi Bt$ qui est l'impulsion dans l'espace BL^2. D'où l'équation :

$$T_Y * H = T_X * \Delta$$

Soit, compte tenu de $T_X * \Delta = T_X$

$$H(t) \in BL_+^2 \quad T_Y * H = T_X \tag{12}$$

III.2 - Détermination de la réponse impulsionnelle à partir de la fonction de Green

Nous nous proposons de calculer la réponse impulsionnelle d'un filtre numérique récursif dont les coefficients $a_m (m=o,\ldots,M)$ et $b_n (n=o,1,\ldots,N)$ sont connus. Pour cela, il faut donc résoudre l'équation de convolution (12).

Or la solution H(t) de (12) peut s'obtenir à partir de la solution G(t) de l'équation suivante :

$$G(t) \in BL_+^2 \quad T_Y * G = \Delta \tag{13}$$

En effet l'équation (12) peut s'écrire compte-tenu de (13)

$$T_Y * H = T_X = T_X * \Delta = T_X * T_Y * G$$

ou :

$$T_Y * (H - T_X * G) = 0 \tag{14}$$

avec : $H(t) \in BL_+^2$ et $G(t) \in BL_+^2$.

Or $G(t) \in BL_+^2 \rightarrow T_Y * G \in BL_+^2$ puisque T_Y est une combinaison linéaire de translations correspondants à des retards. Il en résulte que $(H - T_X * G) \in BL_+^2$ et par suite l'équation (14) est équivalente à une équation aux différences homogène avec des conditions initiales nulles, dont on sait que la solution est identiquement nulle, donc $H - T_X * G \equiv 0$.

soit :

$$H = T_X * G \tag{15}$$

Ce qui nous montre que la solution de l'équation (12) se déduit de la solution de (13) grâce à la relation (15).

G(t) est appelée fonction de Green. Nous allons étudier la résolution de l'équation (13)

III.3 - Calcul de la fonction de Green d'un filtre numérique récursif

Par projection dans la base temporelle, l'équation (13) est équivalente à une équation aux différences avec conditions initiales homogènes, à savoir :

$$\sum_{n=o}^{N} b_n . G_{k-n} = \delta_{k,o} \tag{16}$$

$G_k = 0 \qquad k = -1,-2,\ldots,-N.$

La résolution de cette équation est classique, pour mémoire nous la

rappelons en annexe.

Le résultat obtenu est le suivant :

$$k < 0 \qquad G_k = 0$$

$$k \geqslant 0 \qquad G_k = \sum_{n=1}^{N} \frac{W_{N,n}(-1)}{b_N . W(-1)} \cdot V_n(k) \qquad (17)$$

Les $V_n(k)$ sont les solutions de l'équation algébrique homogène de degré N associée à (16), à savoir :

$$\sum_{n=o}^{N} b_n . z^{N-n} = 0 \qquad (18)$$

- si z_n est racine simple (dans C) de l'équation (18),

$$V_n(k) = z_n^k$$

- si $z_{n'}$ est racine d'ordre p de la même équation :

$$V_{n'+i}(k) = k^i . z_{n'}^k$$

W(-1) est le déterminant NxN suivant :

$$\begin{vmatrix} V_1(-1) & \ldots\ldots & V_N(-1) \\ V_1(-2) & \ldots\ldots & V_N(-2) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ V_1(-N) & \ldots\ldots & V_N(-N) \end{vmatrix}$$

$W_{N,n}(-1)$ est le cofacteur correspondant à l'élément de la ligne N et de la colonne n du déterminant précédent.

III.4 - Expression de la réponse impulsionnelle. Condition de stabilité

III.4.1 - Expression de la réponse impulsionnelle

Nous obtiendrons l'expression de la réponse impulsionnelle, sous la forme des coordonnées H_k de H(t) dans la base temporelle de BL^2 en projetant l'expression (15) dans cette base, il vient :

$$k \in Z \qquad H_k = \sum_{m=o}^{M} a_m . G_{k-m} \qquad (19)$$

où G_k est donné par les relations (17).

III.4.2 - Condition de stabilité

Pour que H(t) déterminé par (17) et (19) soit bien la réponse impulsionnelle d'un filtre numérique récursif, il faut et il suffit que $H(t) \in BL_+^2$. D'après (17) et

(19), on a toujours $\forall k$, $k<o$, $H_k=0$, la condition $H(t) \in BL_+^2$ se réduit donc à : $\sum_{k=o}^{\infty} |H_k|^2 < \infty$, cette dernière condition sera appelée condition de stabilité. Pour étudier cette condition, il faut donc étudier le comportement de H_k pour $k \longrightarrow + \infty$ ($k \in Z$).

D'après l'expression (18), ou bien H_k est nul à partir d'un certain rang, ou bien il se comporte à l'infini comme G_k. Nous étudierons successivement chacun de ces deux cas.

1° - Cas exceptionnel : H_k est nul à partir d'un certain rang

L'équation (12) nous donne par projection sur la base temporelle :

$$k > M, \quad \sum_{n=o}^{N} b_n . H_{k-n} = 0 \quad \text{ou :} \quad H_k = - \sum_{n=1}^{N} b_n . H_{k-n} \qquad (\text{avec } b_o = 1).$$

Cette dernière relation nous montre que si H_k est nul pour N valeurs successives de k supérieures à M-N, il est nul pour toutes les valeurs suivantes de k.

D'après (19), cette condition sera remplie si l'on a :

$$\sum_{m=o}^{M} a_m . G_{k-m} = 0 \quad \text{pour N valeurs successives de k, } k>M-N.$$

Il s'agit d'un système de N équations linéaires et homogènes à M+1 inconnues $(a_o, a_1, \ldots a_M)$. On sait qu'un tel système possède des solutions non nulles si le nombre d'inconnues est supérieur au nombre d'équations, c'est-à-dire :

$$M + 1 > N \qquad \text{ou :} \quad M > N.$$

Si donc $M \geqslant N$, à tout ensemble de coefficients $b_1, b_2, \ldots, b_N$ donnés on peut associer un ou plusieurs ensembles de coefficients $a_o, a_1, \ldots, a_M$ tels que la réponse impulsionnelle du filtre numérique récursif associé ait toutes ses coordonnées nulles à partir d'un certain rang.

Dans ces conditions on a évidemment $\sum_{k=o}^{\infty} |H_k|^2 < \infty$ quel que soit le comportement à l'infini des G_k.

2° - Cas général : H_k se comporte comme G_k lorsque $k \to \infty$

D'après l'expression (17) lorsque $k \to \infty$, G_k se comporte comme $V_n(k)$ qui est de la forme z_n^k ou $k^i . z_n^k$, les z_n étant les racines de l'équation (18). Par suite la condition nécessaire et suffisante pour que la série $|G_k|^2$ et donc aussi la série $|H_k|^2$ convergent est que toutes les racines de l'équation (18) soient de module inférieur à 1.

Remarquons que le 1er cas n'est réalisé que pour des valeurs isolées des coefficients $a_o, a_1, \ldots, a_m$. C'est-à-dire que pour tout écart par rapport à ces valeurs (si faible soit-il) on se retrouve dans le 2ème cas où la stabilité n'est assurée que pour $|z_n|<1$ $(n=1,2,\ldots,N)$.

Le 1er cas n'a donc pas d'existence physique stable, une légère fluctuation des paramètres étant impossible à éviter dans tout appareil physique. D'où la

propriété suivante :

La condition nécessaire et suffisante de stabilité d'un filtre numérique récursif est que les racines de l'équation homogène associée (18) soient de module inférieur à l'unité.

Ces racines sont appelées les pôles du filtre. Elles sont déterminées exclusivement par les coefficients b_n. Ainsi les coefficients a_m n'interviennent pas dans la condition de stabilité.

IV - METHODES D'ETUDE DES FILTRES NUMERIQUES RECURSIFS DANS LE DOMAINE FREQUENTIEL

IV.1 - Gain complexe d'un filtre numérique récursif

IV.1.1 - Expression du gain complexe

Comme les filtres ordinaires, les filtres numériques récursifs peuvent être caractérisés dans le domaine fréquentiel par le gain complexe qui est la transformée de Fourier de la réponse impulsionnelle caractérisant le filtre dans le domaine temporel.

Nous savons que la réponse impulsionnelle H(t) est la solution unique si elle existe de l'équation (12) que nous rappelons ici :

$$H(t) \in BL_+^2 \qquad T_Y * H = T_X \qquad (12)$$

La transformée de Fourier de $T_Y = \sum_{n=o}^{N} b_n . \text{sinc } 2\pi B \left(t - \frac{n}{2B}\right)$

s'écrit : $$t_Y = \Pi_B(\nu) . \sum_{n=o}^{N} b_n . e^{-2\pi i \frac{n}{2B}\nu}$$

De même la transformée de Fourier de T_X est :

$$t_X = \Pi_B(\nu) . \sum_{m=o}^{M} a_m . e^{-2\pi i \frac{m}{2B}\nu}$$

L'équation (12) donne donc par transformée de Fourier :

$|\nu| > B \quad h(\nu) = 0 \quad$ (car, par hypothèse $H(t) \in BL^2$)

$$|\nu| < B \quad h(\nu) . \Pi_B(\nu) . \sum_{n=o}^{N} b_n . e^{-2\pi i \frac{n}{2B}\nu} = \Pi_B(\nu) . \sum_{m=o}^{M} a_m . e^{-2\pi i \frac{m}{2B}\nu}$$

d'où l'expression du gain complexe d'un filtre numérique récursif :

$$h(\nu) = \Pi_B(\nu) . \frac{\sum_{m=o}^{M} a_m . e^{-2\pi i \frac{m}{2B}\nu}}{\sum_{n=o}^{N} b_n . e^{-2\pi i \frac{n}{2B}\nu}} \qquad (20)$$

IV.1.2 - Condition pour qu'une expression du type (20) représente le gain complexe d'un filtre numérique récursif

Nous avons vu au paragraphe 2.III.4.2 que la condition $H(t) \in BL^2$ introduit une contrainte sur les coefficients b_n, que nous avons appelé condition de stabilité, à savoir que les racines de l'équation (18) soient de module inférieur à un.

Il en résulte qu'une expression du type (20) ne correspond au gain complexe d'un filtre numérique récursif que si les coefficients b_n vérifient la condition de stabilité.

Si cette condition n'est pas vérifiée, $H(t) \notin L^2$ et n'a pas de transformée de Fourier dans L^2, l'expression (20) n'est donc plus la transformée de Fourier de la solution de l'équation (12), elle ne représente plus le gain complexe d'un filtre numérique récursif.

IV.2 - Fonction de transfert en Z d'un filtre numérique récursif

IV.2.1 - Expression de la fonction de transfert en Z

Il est d'usage, dans l'expression (20) du gain complexe, de poser $e^{2\pi i \frac{\nu}{2B}} = z$. On a donc $|z| = 1$ et $-\pi \leqslant \mathrm{Arg}\ z \leqslant \pi$ pour $-B \leqslant \nu \leqslant B$. Convenant ainsi de restreindre le domaine de variation de Arg z à l'intervalle $[-\pi, +\pi]$ qui suffit à décrire complètement la réponse du filtre, on peut dans l'expression (20) omettre le facteur $\pi_B(\nu)$ et écrire :

$$H(z) = \frac{\sum_{m=0}^{M} a_m . z^{-m}}{\sum_{n=0}^{N} b_n . z^{-n}} \qquad (21)$$

Lorsque z décrit le plan complexe, H(z) n'est autre que la transformée en z de la réponse impulsionnelle H(t) du filtre numérique récursif. On obtient donc le gain complexe à partir de la fonction de transfert en z lorsque z décrit le cercle unité de $-\pi$ à $+\pi$.

La transformée en z possède de nombreuses propriétés classiques dont nous ne rappellerons ici que la principale : si X(z) est la transformée en z d'une fonction X(t) de BL^2, $z^{-n}.X(z)$ est la transformée en z de $X(t-\frac{n}{2B})$. Cette propriété permet de retrouver directement l'expression (21) à partir de l'équation (8).

IV.2.2 - Condition de stabilité dans le plan des Z

D'après l'expression de la condition de stabilité exposée au paragraphe 2.III.4.2., cette condition se traduit dans le plan des Z de la manière suivante :

La condition de stabilité d'un filtre numérique récursif est que tous les pôles de sa fonction de transfert en z se trouvent à l'intérieur du cercle unité.

IV.3 - Calcul de la norme d'un filtre numérique récursif

Nous regroupons dans ce paragraphe les différentes expressions permettant

le calcul de la norme des filtres numériques et celles s'appliquant plus précisément aux filtres récursifs.

IV.3.1 - Définition

Rappelons que nous avons appelé norme d'un filtre numérique la norme de sa réponse impulsionnelle H(t) dans l'espace BL^2, à savoir :

$$||H|| = \left[\int_R |H(t)|^2 \, dt \right]^{1/2} \qquad (22)$$

IV.3.2 - Expression dans la base temporelle

La base temporelle de BL^2 étant une base de Hilbert, nous avons vu que la norme d'un filtre numérique quelconque peut s'exprimer en fonction des coordonnées dans cette base :

$$||H|| = \left[\sum_{n=-\infty}^{+\infty} |H_n|^2 \right]^{1/2} \qquad (23)$$

en rappelant que $H_n = \frac{1}{\sqrt{2B}} H\left(\frac{n}{2B}\right)$

IV.3.3.- Expression dans le domaine fréquentiel

Cette relation s'applique essentiellement aux filtres récursifs dont nous avons vu l'expression du gain complexe. D'après le théorème de Parseval, on a : $||H(t)|| = ||h(\nu)||$ pour des fonctions de L^2. Ici $h(\nu) \in L^2$ et son support est borné (intervalle [-B,+B]) nous avons donc la relation :

$$||H|| = \left[\int_{-B}^{+B} |h(\nu)|^2 \, d\nu \right]^{1/2} \qquad (24)$$

IV.3.4 - Expression dans le domaine des Z

La relation que nous allons établir s'applique également aux filtres récursifs, pour lesquels on connaît la fonction de transfert en z.

Dans l'intégrale de la fonction (24), posons : $z = e^{2i\pi\frac{\nu}{2B}}$ quand ν varie de -B à +B, z décrit le cercle unité dans le sens positif et l'on a :

$$dz = \frac{1}{2B} \cdot 2\pi i . e^{2\pi i\frac{\nu}{2B}} . d\nu$$

il vient donc l'expression :

$$||H|| = \sqrt{2B} \cdot \left| \frac{1}{2\pi i} \oint H(z).H\left(\frac{1}{z}\right).z^{-1}.dz \right|^{1/2} \qquad (25)$$

Le contour d'intégration étant le cercle unité pris dans le sens positif. En utilisant le théorème des résidus, on peut énoncer la propriété suivante :

Le carré de la norme d'un filtre numérique récursif de fonction de transfert H(z) est égal au produit par 2B de la somme des résidus relatifs aux pôles de l'expression $z^{-1}.H(z).H(\frac{1}{z})$ qui sont intérieurs au cercle unité.

(Le problème des pôles sur le cercle unité ne se pose pas, puisqu'il correspondrait à des filtres instables ou plus précisémment des oscillateurs).

Chapitre 3

SYNTHESE DES FILTRES NUMERIQUES RECURSIFS DANS LE DOMAINE TEMPOREL

I - APPROXIMATION D'UN FILTRE NUMERIQUE PAR UN FILTRE NUMERIQUE RECURSIF

I.1 - Filtre numérique récursif dont la réponse impulsionnelle possède N+1 coordonnées non nulles

Rappelons qu'un filtre numérique récursif est entièrement déterminé par les M+N+1 coefficients a_o, a_1,...,a_m, b_1, b_2,... b_N sous réserve que la condition de stabilité soit vérifiée (voir paragraphe 2.I.2.5 et 2.III.4.2). N est l'ordre du filtre.

Nous allons démontrer la propriété suivante :

On peut toujours trouver un filtre numérique récursif d'ordre N et tel que M=N, dont N+1 coordonnées de la réponse impulsionnelle sont déterminées à l'avance et dont toutes les coordonnées suivantes sont nulles.

En effet, dans le cas M=N l'équation (12) en projection dans la base temporelle nous donne :

$$0 \leqslant k \leqslant M \qquad \sum_{n=o}^{N} b_n . H_{k-n} = a_k \qquad (26)$$

Imposons de plus les conditions :

$$M+1 \leqslant k \leqslant M+N \qquad H_k = 0 \qquad (27)$$

Ce qui permet d'écrire toujours à partir de l'équation (12) :

$$M+1 \leqslant k \leqslant M+N \qquad \sum_{n=k-M}^{N} b_n . H_{k-n} = 0 \qquad (28)$$

Les équations (26) et (28) où les H_k (k=0,1,...,M) sont donnés constituent un système de M+N+1 équations linéaires qui permettent toujours de déterminer (en général de manière unique) les coefficients a_o, a_1,...,a_N, b_1, b_2,..., b_M.

Les équations (26) et (28) étant vérifiées, les conditions (27) sont remplies, le filtre obtenu, possède donc pour k > M, N coordonnées successives de la réponse impulsionnelle qui sont nulles et par suite, en vertu de l'équation (20) toutes les coordonnées suivantes sont également nulles. Nous avons donc :

$$\forall k, \quad k > M \qquad H_k = 0$$

Il en résulte que $||H||^2 = \sum_{n=-\infty}^{+\infty} |H_n|^2 = \sum_{n=0}^{M} |H_n|^2 < \infty$: la condition de stabilité est vérifiée.

Enfin, les coefficients du filtre obtenu vérifiant les équations (26) les N+1 premières coordonnées de la réponse impulsionnelles ont bien les valeurs fixées.

I.2 - Théorème fondamental des filtres numériques récursifs

Nous allons démontrer le théorème suivant, fondamental pour les applications du filtrage numérique récursif :

Tout filtre numérique causal est la limite d'une suite de filtres numériques récursifs pour lesquels M = N.

Cette propriété peut encore s'énoncer :

L'ensemble des réponses impulsionnelles des filtres récursifs tels que M = N est dense dans BL^2_+.

Précisons que la notion de convergence qui intervient dans ces énoncés est celle qui a été définie au paragraphe 1.II.3.2.

Soit H(t) la réponse impulsionnelle d'un filtre numérique causal quelconque dont nous désignerons les coordonnées dans la base temporelle par H_k.

On a $\forall k$, $k < 0$ $H_k = 0$. D'après le théorème du paragraphe précédent, on peut trouver une suite de filtres numériques récursifs, d'ordre N tels que M = N dont les N+1 premières coordonnées de la réponse impulsionnelle $H_N(t)$ sont égales à H_k (k=o,1...,N) et dont toutes les suivantes sont nulles.

Dans ces conditions, nous avons :

$$||H-H_N||^2 = \sum_{k=N+1}^{\infty} |H_k|^2 = ||H||^2 - \sum_{k=o}^{N} |H_k|^2$$

Or, lorsque $N \to \infty$, $\sum_{k=o}^{N} |H_k|^2 \to ||H||^2$, donc $||H-H_N|| \to o$.

En d'autres termes, on peut approcher d'aussi près que l'on veut filtre numérique causal quelconque par un filtre numérique récursif (tel que M=N) à condition de choisir un filtre d'ordre suffisamment élevé.

Remarque : On pourrait constater aisément que toutes les propriétés démontrées pour les filtres récursifs tels que M = N sont a fortiori vraies pour les filtres récursifs tels que M > N (à ordre égal le nombre de coefficients étant plus grand).

II - ETUDE DU PROBLEME DE LA DETERMINATION D'UN FILTRE NUMERIQUE RECURSIF DE REPONSE IMPULSIONNELLE DONNEE

II.1 - Position du problème

Nous savons que la donnée d'une fonction $G(t) \in BL^2$ détermine de manière unique un filtre numérique dont la réponse impulsionnelle est G(t) ; d'après le théorème précédent, il est possible de trouver un filtre numérique récursif dont

la réponse impulsionnelle approche G(t) d'aussi près que l'on veut. Le problème que nous posons ici et que nous tenterons de résoudre pratiquement est moins ambitieux. On sait que G(t) appartenant à BL_+^2 peut être entièrement déterminé par la suite dénombrable de ses coordonnées G_k dans la base temporelle de BL^2. Or $G(t) \in BL_+^2 \rightarrow \lim_{t\to\infty} G(t) = 0$ ou encore $\lim_{n\to\infty} G_n = 0$, nous nous contenterons donc pour définir G(t) d'une suite <u>finie</u> de coordonnées G_k pour $0 \leqslant k \leqslant K$ et nous chercherons à déterminer les coefficients du filtre numérique récursif dont les coordonnées H_k de la réponse impulsionnelle H(t) sont égales ou voisines des G_k pour $0 \leqslant k \leqslant K$ en supposant que les H_k pour $k > K$ seront assez faibles pour que $||H||$ et $||G||$ soient voisins.

II.2 - Mise en équation du problème

Soit un filtre de réponse impulsionnelle G(t), l'équation (8) écrite pour $0 \leqslant k \leqslant K$ nous donne le système d'équations linéaires suivant :

$$
\begin{aligned}
b_o \cdot G_o &= a_o \\
b_1 \cdot G_o + b_o \cdot G_1 &= a_1 \\
&\cdots\cdots\cdots\cdots\cdots \\
b_M \cdot G_o + b_{M+1} \cdot G_1 + \ldots + b_o \cdot G_M &= a_M \qquad (29) \\
b_{M+1} \cdot G_o + b_M \cdot G_1 + \ldots + b_o \cdot G_{M+1} &= 0 \\
&\cdots\cdots\cdots\cdots\cdots \\
b_N \cdot G_{K-N} + b_{N-1} \cdot G_{K-N+1} + \ldots + b_o \cdot G_K &= 0
\end{aligned}
$$

Le problème de trouver un filtre linéaire récursif dont la réponse impulsionnelle est donnée par les K+1 coordonnées G_o, G_1,..., G_K se ramène donc au problème mathématique suivant :

Résoudre un système de K+1 équations linéaires à M+N+1 inconnues (à savoir : $a_o, a_1 \ldots, a_M$, b_1, b_2,..., b_N puisque $b_o = 1$). D'où les 3 cas suivants :

1° - $K < M+N$ c'est-à-dire $K+1 < M+N+1$

Le système possède plus d'inconnues que d'équations, il est indéterminé, les solutions étant constituées par la somme d'une solution particulière et d'une solution du système homogène associé, laquelle appartient à un espace vectoriel à M+N-K dimensions.

Remarquons que ces solutions diffèrent par les valeurs de G_k prises pour $k > K$.

2° - $K = M+N$ c'est-à-dire $K+1 = M+N+1$

Le système possède autant d'équations que d'inconnues, il a donc en général une solution unique.

3° - $K > M+N$ c'est-à-dire $K+1 > M+N+1$

Le système possède plus d'équations que d'inconnues, il est donc en général

impossible, c'est-à-dire qu'il ne possède pas de solution exacte. Le problème que nous étudierons ultérieurement est celui de savoir si ce système possède des solutions approchées, c'est-à-dire des valeurs de a_o, a_1,...,a_M, b_1, b_2,..., b_N donnant une réponse impulsionnelle dont les coordonnées H_k soient voisines de G_k pour $0 \leqslant k \leqslant K$.

Remarquons que pour des raisons technologiques, on a intérêt à rechercher des filtres d'ordre aussi peu élevé que possible, par contre une précision suffisante sur la définition de la réponse impulsionnelle exige une valeur de K assez grande. Il en résulte que pratiquement c'est le cas K > M+N qui retiendra surtout notre attention.

II.3 - Similitude avec le problème de la déconvolution

Le problème général de la déconvolution se pose de la manière suivante. Etant donné un filtre de réponse impulsionnelle H(t) et un signal E(t) à l'entrée qui fournit un signal S(t) à la sortie, on sait que l'on a la relation : S = E*H. Effectuer la déconvolution consiste à essayer de trouver E connaissant S et H ou à essayer de trouver H connaissant S et E.

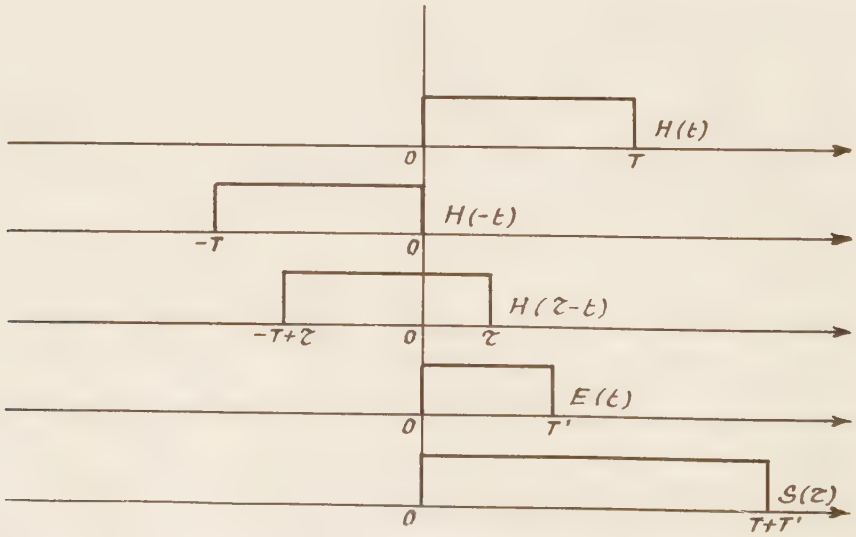
La convolution étant une opération commutative, on voit que les deux questions sont identiques. Dans la suite nous supposerons donc par exemple que S(t) et H(t) sont connus et que E(t) est le signal inconnu.

L'opération générale de convolution S = E*H s'écrit explicitement :

$$S(\tau) = \int_{-\infty}^{+\infty} H(\tau - t).E(t).dt \qquad (30)$$

Cependant dans la pratique, nous travaillons sur des signaux finis qui sont représentés par des fonctions à support borné, ce qui permet de restreindre les bornes de l'intégrale (30).

Si T est la mesure du support de H(t) et T' celle du support de E(t), $S(\tau)$ aura un support de mesure T+T' et l'on peut écrire compte tenu des schémas ci-dessous :



$$\begin{cases} S(\tau) = \displaystyle\int_{\sup\left\{\begin{matrix}\tau-T\\ o\end{matrix}\right.}^{\inf\left\{\begin{matrix}\tau\\ T'\end{matrix}\right.} H(\tau-t).E(t).dt \\ \tau \in [o,\ T+T'] \end{cases} \qquad (31)$$

$$\begin{cases} S(\tau) = 0 \\ \tau \notin [o,\ T+T'] \end{cases}$$

Dans le cas qui nous intéresse, les signaux sont traités numériquement, nous avons donc affaire à des signaux échantillonnés, les intégrales étant remplacées par des sommes.

N et M étant les nombres d'échantillons respectifs de H et E, l'expression (31) s'écrit :

$$\begin{cases} S_i = \displaystyle\sum_{n=\sup\left\{\begin{matrix}i-N\\ o\end{matrix}\right.}^{n=\inf\left\{\begin{matrix}i\\ M\end{matrix}\right.} H_{i-n} \cdot E_n \\ 0 \leqslant i \leqslant M+N \end{cases} \qquad (31')$$

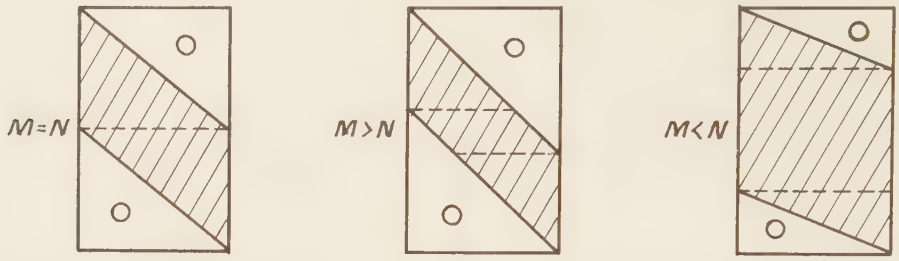
$$\begin{cases} S_i = 0 \\ (i<0)\ U\ (i>M+N) \end{cases}$$

L'expression précédente de S_i peut se mettre sous la forme matricielle suivante :

$$\begin{bmatrix} H_o & & & 0 \\ & H_o & & \\ H_1 & & H_o \cdots & H_o \\ \vdots & H_1 & & \\ \vdots & \vdots & H_1 \cdots & H_1 \\ \vdots & \vdots & \vdots & \vdots \\ H_N & \vdots & \vdots & \vdots \\ & H_N & \vdots & \vdots \\ & & H_N \cdots & H_N \\ & 0 & & \end{bmatrix} \begin{bmatrix} E_o \\ E_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ E_N \end{bmatrix} = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ S_{M+N} \end{bmatrix} \qquad (32)$$

qui constitue un système surdéterminé H.E = S de M+N+1 équations à M+1 inconnues E_o, E_1,...,E_M.

Notons que la matrice H qui est une matrice à M+N+1 lignes et M+1 colonnes prend l'allure indiquée sur le schéma, selon les valeurs relatives de M et N :



En définitive, nous constatons que nous sommes ramenés à un problème mathématique identique à celui de la détermination des coefficients d'un filtre récursif dans le 3ème cas envisagé au paragraphe 2 (K > M+N).

C'est pourquoi nous étudierons dans le paragraphe suivant quelques éléments permettant d'aborder ce problème mathématique.

Remarque : On peut dire aussi, du point de vue physique que le problème de la détermination des coefficients d'un filtre récursif revient à trouver les caractéristiques d'un système dont on connaît le signal d'entrée (impulsion unitaire) et la sortie (réponse impulsionnelle) et en ce sens il s'agit bien d'un problème de déconvolution.

III - NOTIONS MATHEMATIQUES RELATIVES AUX SYSTEMES LINEAIRES SURDETERMINES

III.1 - Généralités

Considérons donc le système linéaire AX = B dans lequel :

- A est une matrice mxn avec m>n
- X est une matrice unicolonne nx1
- B est une matrice unicolonne mx1

On sait que ce système est en général impossible, mais on peut se poser la question de savoir s'il possède des solutions approchées telles que nous allons les définir.

Considérons une matrice Y(nx1) et écrivons :

$$AY = B + \varepsilon$$

où ε est une matrice unicolonne mx1.

La démarche intuitive consiste à essayer de trouver une valeur de Y telle que B + ε soit "voisine" de B, c'est-à-dire ε "proche" de zéro.

Cette notion de proximité trouve sa formulation mathématique rigoureuse dans la notion de norme.

L'équation AX = B peut être considérée comme l'expression du problème suivant : les espaces vectoriels E_n (n dimensions) et F_m (m dimensions) étant rapportés à des bases déterminées, la matrice A représente dans ces bases une application linéaire f de E_n dans F_m. Existe-t-il dans E_n un vecteur X dont l'image par f est

un vecteur B de F_m ?.

L'image par f de E_n étant un sous espace vectoriel de F_m dont la dimension est inférieure ou égale à n, le problème n'a de solution exacte que si B appartient à ce sous espace vectoriel, ce qui n'est pas le cas en général.

Introduisons des normes soit N_1 dans E_n et N_2 dans F_m.

Nous nous proposons de répondre alors à deux questions :

1° - Posant $AY = B+\varepsilon$, peut-on trouver $Y = Y_o$ tel que $N_2(\varepsilon_o)$ soit minimal.

2° - Quel rapport existe-t-il entre $N_2(\varepsilon-\varepsilon_o)$ et $N_1(Y-Y_o)$, autrement dit une petite variation de $N_2(\varepsilon)$ entraîne-t-elle une petite variation de $N_1(Y)$. Cette question a pour cas particulier la question suivante qui est d'un grand intérêt pour les applications : Dans le cas où le système possède une solution exacte X $[N_2(\varepsilon_X)=0]$, le fait de trouver une solution approchée Y pour laquelle $N_2(\varepsilon_Y)$ est voisin de 0, implique-t-il que Y est proche de X c'est-à-dire que $N_1(Y-X) \simeq 0$?

C'est le problème de la stabilité des systèmes linéaires, il est lié à la notion de conditionnement des matrices dont nous allons donner un aperçu au paragraphe suivant.

III.2 - Solution approchée au sens de la minimisation de la norme euclidienne : Pseudo-inverses [14]

Les études que nous avons faites portent exclusivement sur l'utilisation des normes euclidiennes en base orthonormé, à savoir :

$$N_2(\varepsilon) = \left[\sum_{i=1}^{m} \varepsilon_i^2 \right]^{1/2}$$

$$N_1(Y) = \left[\sum_{i=1}^{n} y_i^2 \right]^{1/2}$$

Le principal résultat de cette étude répond à la 1ère question du paragraphe précédent, il s'énonce comme suit :

Etant donné le système $AY = B+\varepsilon$ où A est une matrice avec $m > n$ il existe une solution Y_o en général unique qui minimise la norme euclidienne de ε. Cette solution est donnée par :

$$Y_o = A^{(-1)}.B \qquad (33)$$

où $A^{(-1)}$ est la pseudo-inverse de la matrice A.

Précisons que la pseudo-inverse $A^{(-1)}$ d'une matrice rectangulaire $A_{m,n}$ (m lignes, n colonnes) est la matrice rectangulaire $A^{(-1)}_{n,m}$ (n lignes, m colonnes) satisfaisant aux relations suivantes : 6)

---

6) La notation * indiquant la transposée- conjuguée d'une matrice.

$$\begin{cases} A.A^{(-1)}.A = A \\ A^{(-1)}.A.A^{(-1)} = A^{(-1)} \\ (A^*)^{(-1)} = (A^{(-1)})^* \\ A.A^{(-1)} = [A.A^{(-1)}]^* \end{cases}$$

On démontre que la matrice $A^{(-1)}$ ainsi définie est unique.

Dans le cas où la matrice A est de rang maximal, c'est-à-dire si la matrice carré A*.A (pour m>n) ou AA* (pour m<n) est régulière, le pseudo inverse $A^{(-1)}$ est donnée par :

$$A_{n,m}^{(-1)} = (A^*A)^{-1} \cdot A^* \qquad (34)$$

(cas m > n)

$$A_{n,m}^{(-1)} = A^* \cdot (A.A^*)^{-1} \qquad (34')$$

(cas m < n)

III.3 - Notion de conditionnement de la matrice d'un système linéaire

Cette notion nous sera utile pour essayer d'apporter une réponse à la 2ème question du paragraphe 3.1.1.

Considérons l'exemple suivant :

1° Le système $\begin{cases} 2x + 3y = 5 \\ 2,01x + 2,98y = 4,99 \end{cases}$ a pour solution $\begin{cases} x = 1 \\ y = 1 \end{cases}$

le système $\begin{cases} 2x + 3y = 5 \\ 2x + 3,01y = 4,98 \end{cases}$ déduit du précédent

par de légères perturbations sur les coefficients a pour solution :

$$\begin{cases} x = 5,5 \\ y = -2 \end{cases}$$

2° Le système $\begin{cases} 5x - 2y = 3 \\ 6x + 15y = 21 \end{cases}$ a pour solution $\begin{cases} x = 1 \\ y = 1 \end{cases}$

la système $\begin{cases} 5x - 2y = 3 \\ 6x - 15,01y = 20,98 \end{cases}$ déduit du précédent

pas de légères perturbations sur les coefficients a pour solution :

$$\begin{cases} x = 0,999 \\ y = 0,998 \end{cases}$$

Nous constatons que dans le 1er système une légère perturbation sur les coefficients conduit à une solution totalement différente de la solution initiale tandis que dans le 2ème système une même perturbation conduit à une solution très voisine de la solution initiale.

Nous dirons que dans le 1er cas le système est mal conditionné tandis que

dans le 2ème il est bien conditionné.

Plus généralement un système d'équations linéaires AX = B (A matrice carrée) a un conditionnement qui dépend de la matrice A.

On cherche à caractériser ce conditionnement d'une matrice carrée A par un nombre réel positif.

Plusieurs définitions existent. La plus courante est :

$$\frac{\inf_i |\lambda_i|}{\sup_i |\lambda_i|}$$

λ_i étant les valeurs propres. A. GASTINEL [10], propose une autre définition appelée conditionnement normalisé dans laquelle le nombre de conditionnement est compris entre 0 et 1.

Géométriquement le conditionnement peut s'interpréter de la manière suivante: la solution d'un système de n équations linéaires à n inconnues peut être considérée comme les coordonnées de l'intersection des n hyperplans définis par les n équations dans un espace affine rapporté à un repère arbitraire. Si l'on dote cet espace d'une structure euclidienne, nous pourrons parler d'angle de ces hyperplans et nous dirons qu'un système mal conditionné correspond à un ensemble d'hyperplans faisant entre eux des angles très faibles. En effet, dans ce cas on conçoit qu'un léger déplacement de l'un des hyperplans déplace sensiblement leur intersection. Les systèmes les mieux conditionnés seront donc ceux correspondants à des matrices orthogonales (hyperplans orthogonaux).

IV - APPLICATION A LA DETERMINATION DES COEFFICIENTS A PARTIR DE K COORDONNEES DE LA REPONSE IMPULSIONNELLE

IV.1 - Méthodes utilisées

IV.1.1 - Méthodes de Burrus et Parks [2]

Cette méthode consiste à scinder en deux parties le système d'équations (29) :

1° Les K-M dernières équations peuvent s'écrire (avec $b_o = 1$) :

$$b_{M+1} \cdot G_o + b_M \cdot G_1 + \ldots + b_1 \cdot G_M = - G_{M+1}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$b_N \cdot G_{K-N} + b_{N-1} \cdot G_{K-N+1} + \ldots + b_1 \cdot G_{K-1} = - G_K$$

Elles constituent un système en général surdéterminé de K-M équations à N inconnues ($b_1, b_2, \ldots, b_N$) dont on peut trouver une solution approchée au sens de la minimisation de la norme euclidienne comme nous l'avons vu en III.2.

2° Une fois $b_1, b_2, \ldots b_N$ déterminés, les M+1 premières équations du système (29) constituent un système à M+1 inconnues $a_o, a_1, \ldots, a_M$ qui, en général, possède une solution unique déterminant sans ambiguité les coefficients $a_o, a_1, \ldots, a_M$. Les valeurs ainsi trouvées constituent les coefficients du filtre cherché, sous

réserve que la condition de stabilité soit vérifiée.

Cette méthode présente l'inconvénient de répartir sur K-M points seulement l'erreur sur la réponse impulsionnelle obtenue. En effet, le calcul des coefficients $a_o, a_1, \ldots, a_M$ étant exact une fois les $b_1, \ldots, b_N$ choisis, on constate que les M+1 premiers points de la réponse impulsionnelle obtenus sont identiques à ceux du modèle, tout l'écart entre le résultat du calcul et le modèle proposé étant concentré sur les K-M derniers points.

IV.1.2 - Méthode directe

Elle consiste à reprendre directement le système (29) en le mettant sous la forme suivante (avec $b_o = 1$ et donc $a_o = G_o$)

$$b_1 . G_o - a_1 = - G_1$$

$$b_2 . G_o + b_1 . G_1 - a_2 = - G_2$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$b_M . G_o + \ldots + b_1 . G_{M-1} - a_M = - G_M$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$b_N . G_{K-N} + \ldots + b_1 . G_{K-1} = - G_k$$

On a ainsi un système de K équations à M+N inconnues que l'on peut encore écrire sous forme matricielle :

$$\begin{bmatrix} G_o & 0 & 0 & \cdots & 0 & -1 & \cdots & 0 \\ G_1 & G_o & 0 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \cdots & & & & & & & & \\ G_{M-1} & G_{M-2} & \cdots & G_o & 0 & 0 & 0 & \cdots & -1 \\ \cdots & & & & & & & & \\ G_{N-1} & G_{N-2} & \cdots & G_o & 0 & 0 & \cdots & 0 \\ \cdots & & & & & & & & \\ G_{K-1} & G_{K-2} & \cdots & G_{K-N} & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \\ a_1 \\ \vdots \\ a_M \end{bmatrix} = - \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_K \end{bmatrix}$$

Pour résoudre ce système, nous avons utilisé deux méthodes :

1° - En utilisant la formule (33) qui nécessite le calcul de la pseudo-inverse selon la relation (34).

2° - En utilisant une méthode due à Golub [12] qui présente l'avantage d'éviter le calcul de la pseudo-inverse de la matrice particulièrement délicat à cause du mauvais conditionnement de la matrice carrée A*.A qu'il faut inverser.

IV.2 - Résultats obtenus

A partir d'un modèle donné, avec K=40 nous avons à titre expérimental calculé les coefficients a_i et b_i ainsi que la réponse impulsionnelle du filtre

ainsi obtenu pour des valeurs de M en N variant de M=N=6 jusqu'à M=N=20 (K=M+N c'est-à-dire la solution exacte). On constatera que l'écart avec le modèle diminue jusqu'à s'annuler pour K = M+N.

Nous avons chiffré cet écart de la manière suivante : les K coordonnées G_1, $G_2,\ldots,G_K$ de la réponse impulsionnelle modèle constituent les coordonnées d'un vecteur $\vec{G}$ d'un espace vectoriel E_K de dimension K, de même les K coordonnées H_1, $H_2,\ldots,H_K$ de la réponse réelle du filtre obtenues pour M et N donnés constituent les coordonnées d'un vecteur $\vec{H}$ du même espace vectoriel. On peut alors caractériser l'écart entre le filtre obtenu et le modèle souhaité par la distance relative entre les vecteurs $\vec{H}$ et $\vec{G}$ prise au sens de la norme Euclidienne dans E_K soit :

$$d = \frac{||\vec{H}-\vec{G}||}{||\vec{G}||} = \left[\frac{\sum_{i=1}^{K} |H_i-G_i|^2}{\sum_{i=1}^{K} |G_i|^2} \right]^{1/2} \qquad (37)$$

On trouvera dans le tableau ci-dessous les valeurs de d en fonction de M et N calculées par les deux méthodes précédemment décrites, on trouvera également les courbes correspondantes obtenues par la méthode de Burrus et Parks à l'annexe IV.

valeur de M = N	Méthode de Burrus et Parks	Méthode directe
6	0,413	0,415
8	0,312	0,320
10	0,276	0,280
12	0,138	0,138
14	0,095	0,095
16	0,062	0,063
18	0,031	0,041
20	0,000	0,000

On voit d'après ce tableau que les résultats des deux méthodes sont sensiblement équivalents.

Chapitre 4

PROBLEMES RELATIFS A LA REALISATION DES FILTRES NUMERIQUES RECURSIFS

INTRODUCTION

Jusqu'à présent nous n'avons considéré les filtres numériques que comme des systèmes échantillonnés. Dans ces conditions le traitement effectué par ces filtres était supposé exact, pourvu qu'il s'applique à des signaux de spectre borné échantillonnés à une cadence convenable. Or en fait, dans un filtre réel cablé, comme dans tout calculateur numérique, la valeur des échantillons traités est représentée par un nombre binaire, dont le nombre de bits est nécessairement fini. Il en résulte que les valeurs des échantillons du signal traité sont en fait quantifiées, cette quantification introduisant une valeur dans le processus de filtrage, erreur que nous allons étudier.

I - ANALYSE QUALITATIVE DES SOURCES D'ERREUR DANS UN FILTRE NUMERIQUE RECURSIF CABLE

Les effets de la quantification introduisent dans le traitement numérique effectué par le filtre des erreurs que l'on peut classer en trois groupes dont la contribution à l'erreur globale sur le signal de sortie est différente. Nous verrons ensuite dans quelle mesure ces erreurs peuvent engendrer des instabilités.

I.1 - Erreur due à la quantification du signal d'entrée

C'est l'erreur la plus immédiate à déceler : lorsque le signal d'entrée est fourni sous forme analogique il doit passer par l'intermédiaire d'un convertisseur analogique digital (CAD) situé à l'entrée de l'appareil et qui a pour but de l'échantillonner puis d'en quantifier les échantillons. Cette quantification introduit une erreur due à l'écart entre les valeurs vraies des échantillons et les valeurs quantifiées prises en compte. Il en résulte donc une erreur sur les valeurs des échantillons de sortie.

Rappelons qu'il existe trois types de quantification selon le CAD employé :

- la quantification d'arrondi : la valeur prise en compte correspond au niveau de quantification le plus proche de la valeur exacte.

- la quantification de seuil : la valeur prise en compte correspond au niveau de quantification immédiatement inférieur à la valeur exacte.

- la quantification de seuil en amplitude : la valeur prise en compte correspond au niveau de quantification immédiatement inférieur si cette valeur est positive, et immédiatement supérieur si cette valeur est négative.

I.2 - Erreur due à la quantification des paramètres

Les paramètres $a_m (m=0,\ldots,M)$ et $b_n (n=1,\ldots,N)$ étant affichés numériquement, ils sont également quantifiés ce qui entraîne que leur valeur réelle diffère en général légèrement de leur valeur exacte choisie en raison de cette erreur de quantification.

I.3 - Erreur d'arrondi dans les multiplications

La précision du multiplicateur (nombre de bits des mots sur lesquels il opère) étant constante il est nécessaire de tronquer les résultats lorsqu'ils sont destinés à être réutilisés par le multiplieur. Ainsi le résultat du produit de deux mots de N bits est en général un mot de 2N bits dont il faudra rejeter les N bits les moins significatifs si l'on désire le réemployer dans une autre multiplication. Cette opération n'est autre qu'une nouvelle quantification qui cette fois n'a plus lieu sur des valeurs d'entrée, mais sur des valeurs représentant des intermédiaires de calcul.

I.4 - Erreurs et stabilité

I.4.1 - Erreurs n'affectant pas la stabilité

Nous avons vu que lorsque la condition de stabilité est vérifiée, le filtre correspondant est stable pour tous les signaux $X(t) \in BL^2$ que l'on présente à l'entrée, c'est-à-dire que la sortie correspondante est toujours un signal $Y(t) \in BL^2$, il n'y a donc en sortie ni oscillations, ni dérive correspondant à des signaux indéfiniment croissants.

Or, la quantification du signal d'entrée correspond en fait au remplacement du vrai signal d'entrée par un autre signal qui appartient également à BL^2, il en résulte que l'erreur due à la quantification du signal d'entrée ne peut en aucun cas être génératrice d'instabilité.

I.4.2 - Erreurs pouvant engendrer des instabilités

Au contraire, la quantification des coefficients du filtre en modifiant la valeur exacte des coefficients déplace la position réelle de pôles qui peuvent ainsi sortir du cercle unité et donc rendre le filtre instable. Il en va de même de l'erreur d'arrondi dans les multiplications qui en modifiant le résultat des calculs sera équivalent à un déplacement des pôles pouvant se retrouver à l'extérieur du cercle unité. Toutefois dans ce cas l'erreur affectant le produit échantillon du signal par coefficient du filtre, l'instabilité peut n'apparaître que pour certains signaux.

I.4.3 - Degré de stabilité d'un filtre numérique récursif

L'instabilité apparaît dès que les erreurs sont suffisantes pour provoquer un déplacement des pôles tels qu'ils sortent du cercle unité. Il en résulte qu'un filtre sera d'autant plus stable que les perturbations nécessaires pour que les pôles sortent du cercle unité seront importantes.

Ce critère dépend lui-même de deux éléments :

- la proximité des pôles par rapport au cercle unité en l'absence de toute perturbation (filtre idéal) ce facteur dépend de la nature du filtre (ainsi les filtres à bande étroite ont en général des pôles proches du cercle unité et sont donc sujets aux instabilités),

- l'importance du déplacement des pôles correspondant à une perturbation donnée sur les coefficients, ce facteur dépend de la structure et de l'ordre du filtre.

Les risques d'instabilités constituent le principal écueil dans la construction des filtres numériques récursifs ; l'étude qui va suivre est indispensable à effectuer avant d'entamer la construction d'un tel filtre, elle a pour ambition de déterminer la structure correspondant à la stabilité maximale pour un filtre de nature donnée. Nous passerons d'abord en revu les différentes structures avant d'entreprendre l'étude des erreurs de quantification et de la stabilité.

II - DIFFERENTES SRTUCTURES DE FILTRES NUMERIQUES RECURSIFS

Nous avons vu qu'un filtre numérique récursif est caractérisé par la donnée des M+N+1 coefficients a_m(m=0 à M) et b_n(n=1 à N) qui nous permet de former soit l'équation de récurrence du filtre, soit sa fonction de transfert en Z. Selon les structures que nous voulons faire apparaître, nous utiliserons l'une ou l'autre de ces deux représentations. D'autre part, la propriété fondamentale établie au paragraphe 3.I.2 nous autorise à nous borner au cas où $M \leqslant N$.

II.1 - Structure directe

L'équation (8) nous conduit à la structure dite directe de la figure 4.1

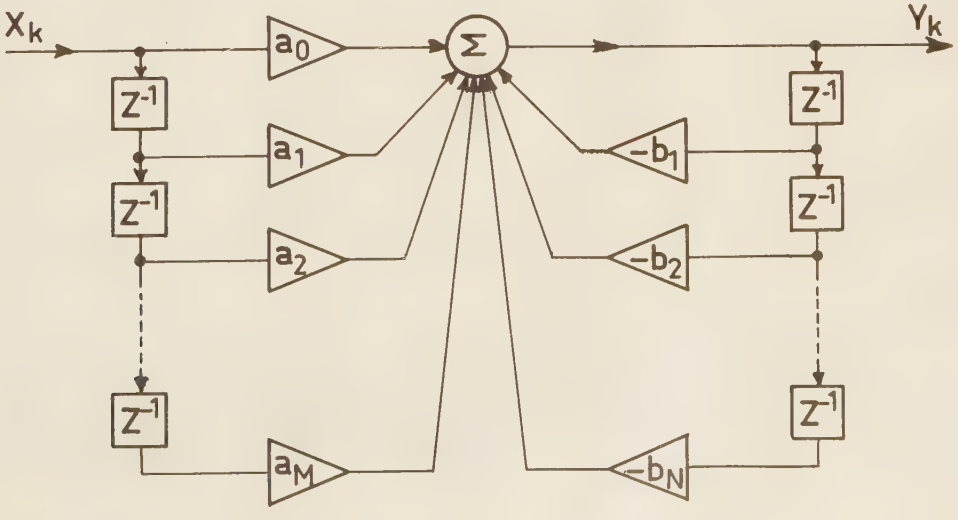


Figure 4.1

II.2 - Structure canonique

La fonction de transfert en Z s'écrit : $H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{m=o}^{M} a_m . z^{-m}}{\sum_{n=o}^{N} b_n . z^{-n}}$

En posant $T(z) = \frac{X(z)}{\sum_{n=o}^{N} b_n . z^{-1}}$ avec $b_o = 1$, il vient :

$$T(z) = X(z) - \sum_{n=1}^{N} b_n . z^{-1} . T(z) \qquad (38)$$

et d'après l'expression de H(z) :

$$Y(z) = \sum_{m=o}^{M} a_m . z^{-m} . T(z) \qquad (39)$$

Les équations (38) et (39) conduisent dans le domaine temporel à deux équations de récurence portant sur les suites d'échantillons $\{X_k\}$ et $\{Y_k\}$ des signaux X(t) et Y(t) de BL_+^2 ainsi que sur la suite des échantillons T_k d'un signal intermédiaire T(t) soit :

$$T_k = X_k - \sum_{n=1}^{N} b_n . T_{k-n} \qquad (40)$$

$$Y_k = \sum_{m=o}^{M} a_m . T_{k-m} \qquad (41)$$

Ces équations conduisent à la structure de la figure 4.2 dite structure canonique.

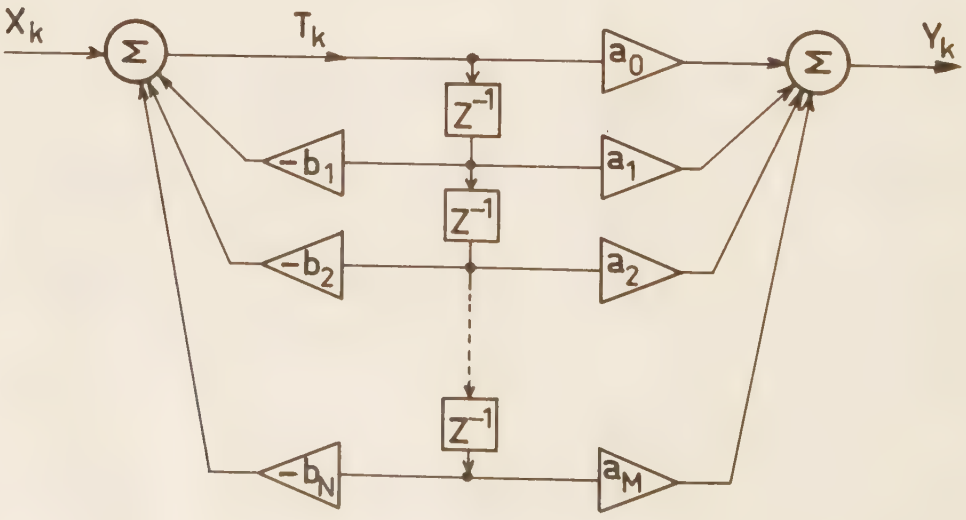


Figure 4.2

II.3 - Structure à éléments en cascade

La fonction de transfert en Z étant le quotient d'un polynôme en z de degré M par un polynôme en z de degré N (et dont le coefficient b_o du terme de plus haut degré est pris égal à 1) elle peut se mettre sous la forme d'un produit :

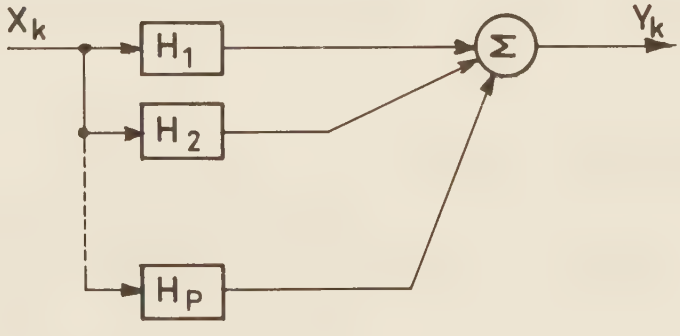
$$H(z) = \prod_{i=1}^{P} H_i(z)$$

avec : $H_i(z) = \dfrac{\alpha+\beta.z^{-1}+\gamma.z^{-2}}{1+\delta.z^{-1}.\varepsilon.z^{-2}}$ $\qquad M \leqslant N, \quad P = \text{Ent}\left[\dfrac{N+1}{2}\right]$

Les pôles des fractions rationnelles $H_i(z)$ sont ceux de H(z), si donc H(z) représente un filtre stable, tous les $H_i(z)$ (dont les coefficients sont toujours réels puisqu'on s'est limité à une décomposition en facteurs du second degré) représentent des filtres récursifs du 2ème ordre stables. On obtient donc la structure à base d'éléments du second ordre associés en cascade représentée par la figure 4.3 :

Figure 4.3

Figure 4.4



II.4 - Structure à éléments en parallèle

Dans le cas M ⩽ N que nous envisageons, la fonction de transfert H(z) est une fraction rationnelle en z qui peut se décomposer en éléments simples sur le corps des réels sous la forme suivante : 7)

$$H(z) = \sum_{j=o}^{P} H_j(z) \qquad P = \text{Ent}\left[\frac{N+1}{2}\right]$$

avec $H_j(z) = \dfrac{A_j+B_j.z^{-1}}{1+C_j.z^{-1}+D_j.z^{-2}}$ $\qquad (B_o=C_o=D_o=0)$

7) La présence de pôles doubles conduirait à une forme légèrement différente par l'introduction d'éléments d'ordre supérieur à deux.

Comme précédemment on remarque que les pôles des $H_j(z)$ sont ceux de $H(z)$, si donc le filtre représenté par $H(z)$ est stable il en va de même des filtres récursifs du second ordre représentés par les $H_j(z)$.

On obtient donc la structure à base d'éléments du second ordre en parallèle représentée par la figure 4.4.

II.5 - Remarque

Dans les deux derniers cas, les éléments du 2ème ordre peuvent dégénérer en éléments du 1er ordre (ou même comporter un élément d'ordre 0). Les éléments qui sont réellement du 2ème ordre peuvent eux-même avoir soit la structure directe, soit la structure canonique.

III - ETUDE DE L'ERREUR DE QUANTIFICATION

III.1 - Généralités

Nous avons vu au paragraphe précédent que toutes les erreurs avaient pour origine l'opération de quantification inévitable dans tout traitement numérique, c'est pourquoi nous allons rappeler les principaux résultats théoriques relatifs à la quantification ; toutefois, l'application de ces résultats ne nous dispensera pas d'un travail de simulation sur calculateur qui seul permet de vérifier la valeur des hypothèses faites dans les calculs théoriques, ainsi que l'affirment différents auteurs. [8)]

III.2 - Quantification d'un signal certain

Soit $X(t)$ un signal de BL^2. A l'instant t_o la quantification associe à la valeur vraie, $X_o = X(t_o)$ la valeur quantifiée $\hat{X}_o$. La transformation $X_o \to \hat{X}_o$ dépend du type de quantification choisie mais dans tous les cas il s'agit d'une transformation non linéaire, ainsi par exemple, dans le cas de la quantification de seuil on a :

$$\hat{X}_o = q.\left[\operatorname{sgn} X_o . \operatorname{Ent}\left(\frac{|X_o|}{q}\right) - \frac{1-\operatorname{sgn} X_o}{2}\right]$$

---

8) "L'introduction de la quantification dans les filtres entraîne des relations non-linéaires qui, sauf dans un certain nombre de cas simples, sont difficiles, sinon impossibles à traiter de manière rigoureuse. Heureusement lorsque le pas de quantification est petit par rapport aux valeurs du signal et des paramètres, la théorie se simplifie et de nombreux résultats utiles peuvent être obtenus et vérifiés expérimentalement". [11] chapitre 4, page 98.

"De l'avis de l'auteur, le moyen le plus sûr et peut-être le plus rapide à long terme, d'analyser l'effet des erreurs de quantification est d'utiliser une simulation sur calculateur" [13] chapitre 5, page 27.

avec : q = pas de quantification
Ent = fonction partie entière
$\operatorname{sgn} X_o$ = fonction signe : $\operatorname{sgn} X_o = +1$ si $X_o \geq 0$
$\operatorname{sgn} X_o = -1$ si $X_o < 0$

La réponse d'un filtre linéaire $\mathcal{H}$, du fait de la quantification sera $\mathcal{H}[\hat{X}(t)]$ au lieu de $\mathcal{H}[X(t)]$, mais la non linéarité de la correspondance $X(t) \to \hat{X}(t)$ ne nous permet pas de pousser le calcul plus avant, c'est pourquoi il faut renoncer à la méthode exacte dans le calcul des erreurs de quantification. Nous allons donc utiliser une méthode statistique.

III.3 - Etude statistique de la quantification

III.3.1 - Principe de la méthode

Nous considérerons ici X(t) comme un signal aléatoire stationnaire et ergodique dont on connaît les moments du premier et du second ordre. Ceci est en fait équivalent à la donnée d'un signal certain $X_c(t)$ ($t \in]-\infty,+\infty[$) qui serait l'une des réalisations du signal aléatoire X(t) et dont on ne connaîtrait que la moyenne

temporelle : $m = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} X_c(t).dt$ et la fonction de corrélation

$\Gamma(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} X_c(t).X_c(t-\tau).dt$ [9] . Nous chercherons à évaluer l'influence

sur m et $\Gamma(\tau)$ de la quantification du signal X(t) et nous verrons que de ces calculs nous autorise moyennant certaines approximations, à remplacer l'opération de quantification par une autre statistiquement équivalente (addition d'un bruit blanc) mais présentant l'avantage d'être linéaire ce qui nous permettra d'évaluer les perturbations crées après le passage dans un filtre linéaire.

III.3.2 - Quantification d'une variable aléatoire à une dimension

Soit X une variable aléatoire à une dimension, de fonction caractéristique $\phi(u)$. A X on associe la variable aléatoire quantifiée $\hat{X}$ de la manière suivante :

$$X = n.q \quad \text{si} \quad \left(n - \tfrac{1}{2}\right).q < X \leq \left(n + \tfrac{1}{2}\right).q$$

(quantification d'arrondi au pas q).

Posant $\mathcal{R}e\,[\phi(u)]$ = partie réelle de $\phi(u)$
$\mathcal{I}m\,[\phi(u)]$ = partie imaginaire de $\phi(u)$

$$\phi'(u) = \frac{d\phi(u)}{du}$$

---

9) X(t) étant stationnaire, $X_c(t) \in L^2$ d'où la nécessité des coefficients $\frac{1}{2T}$.

On démontre (voir annexe III) les expressions suivantes des deux premiers moments de $\hat{X}$:

$$E\{\hat{X}\} = E\{X\} + \frac{q}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \cdot \mathfrak{Im}\left[\phi\left(\frac{2k\pi}{q}\right)\right] \qquad (42)$$

$$E\{\hat{X}^2\} = E\{X^2\} + \frac{q^2}{12} + \frac{q^2}{\pi^2} \cdot \sum_{k=1} \frac{(-1)^k}{k^2} \cdot \mathfrak{Re}\left[\phi\left(\frac{2k\pi}{q}\right)\right] - \frac{2q}{\pi} \cdot \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \cdot \mathfrak{Re}\left[\phi'\left(\frac{2k\pi}{q}\right)\right] \qquad (43)$$

Ces expressions sont assez complexes, cependant sous certaines conditions elles peuvent se simplifier considérablement :

1° - Si $\mathfrak{Im}\left[\phi\left(\frac{2k\pi}{q}\right)\right] = 0$ pour tout k (ce qui se produit en particulier si la densité de probabilité de X est paire, $\phi(u)$ étant alors réelle), la première relation devient :

$$E\{\hat{X}\} = E\{X\} \qquad (44)$$

2° - Si $\forall k, \mathfrak{Re}\left[\phi\left(\frac{2k\pi}{q}\right)\right] = 0$ et $\mathfrak{Re}\left[\phi'\left(\frac{2k\pi}{q}\right)\right] = 0$ l'expression (43) devient :

$$E\{\hat{X}^2\} = E\{X^2\} + \frac{q^2}{12} \qquad (45)$$

Cette condition est notamment réalisée :

a) - d'une manière rigoureuse si $\phi(u)$ est une fonction à support borné par la quantité Q telle que $Q < \frac{2\pi}{q}$ ou $q < \frac{2\pi}{Q}$: ceci nécessite (d'après les propriétés de la transformée de Fourier) que la densité de probabilité p(x) soit indéfiniment dérivable et à dérivées bornées [10].

b) - d'une manière approximative dans le cas d'un processus Gaussien si q soit suffisamment petit par rapport à σ afin que $\phi\left(\frac{2\pi}{q}\right)$ soit négligeable devant 1 et $\phi'\left(\frac{2\pi}{q}\right)$ négligeable devant q. [11]

III.3.3 - Quantification d'une variable aléatoire à deux dimensions

Soit X une variable aléatoire à deux dimensions de composantes X_1 et X_2 et de fonction caractéristique $\phi_{12}(u_1,u_2)$. A X on associe la variable aléatoire $\hat{X}$ de la manière suivante :

---

10) D'après le théorème de Bernstein la dérivée $n^{ème}$ doit être bornée par $\frac{M}{q^n}$ (M borne supérieure de p(x)).

11) pour $q = \frac{2\pi}{3}.\sigma$ on a déjà : $\left|\phi\left(\frac{2\pi}{q}\right)\right| = 0{,}01$ et $\frac{1}{q} \cdot \left|\phi'\left(\frac{2\pi}{q}\right)\right| = 0{,}05$.

$$\hat{X}_1 = n.q \qquad (n-\tfrac{1}{2}).q < X_1 \leqslant (n+\tfrac{1}{2}).q$$

$$\hat{X}_2 = m.q \qquad (m-\tfrac{1}{2}).q < X_2 \leqslant (m+\tfrac{1}{2}).q$$

posant $\mathcal{R}e\left[\phi(u_1,u_2)\right]$ = partie réelle de $\phi(u_1,u_2)$

$$\phi'_1(u_1,u_2) = \frac{\partial\phi(u_1,u_2)}{\partial u_1}$$

$$\phi'_2(u_1,u_2) = \frac{\partial\phi(u_1,u_2)}{\partial u_2}$$

On démontre (voir annexe III) l'expression suivante du moment mixte de la variable aléatoire quantifiée :

$$E\{\hat{X}_1,\hat{X}_2\} = E\{X_1.X_2\} + \frac{q^2}{2\pi^2}\cdot\sum_{\substack{k_2=-\infty\\k_2\neq o}}^{+\infty}\sum_{\substack{k_1=-\infty\\k_1\neq o}}^{+\infty}\frac{(-1)^{k_1+k_2}}{k_1.k_2}\cdot\phi\left(\frac{2k_1\pi}{q},\frac{2k_2\pi}{q}\right)$$

$$-\frac{q}{2\pi}\cdot\sum_{\substack{k_1=-\infty\\k_1\neq o}}^{+\infty}\frac{(-1)^{k_1}}{k_1}\cdot\phi'_2\left(\frac{2k_1\pi}{q},0\right)-\frac{q}{2}\sum_{\substack{k_2=-\infty\\k_2\neq o}}^{+\infty}\frac{(-1)^{k_2}}{k_2}\,\phi'_1\left(0,\frac{2k_2\pi}{q}\right) \qquad (46)$$

Moyennant des hypothèses analogues à celles du paragraphe précédent nous envisageons le cas où :

$$\forall k_1,\ \forall k_2 \quad \phi_{12}\left(\frac{2k_1\pi}{q},\frac{2k_2\pi}{q}\right) = 0 \qquad \text{ou négligeable devant 1}$$

$$\forall k_1,\ \forall k_2 \quad \phi'_1\left(0,\frac{2k_2\pi}{q}\right) = \phi'_2\left(\frac{2k_1\pi}{q},0\right) = 0 \qquad \text{ou négligeable devant } \frac{1}{q}$$

dans ces conditions la relation (46) prend la forme très simple :

$$E\{\hat{X}_1.\hat{X}_2\} = E\{X_1.X_2\}$$

III.3.4 - Application à la quantification d'un signal aléatoire

Soit X(t) un processus aléatoire stationnaire et ergodique dont la densité spectrale énergétique est à support B borné. Il est donc échantillonnable à la cadence de Shannon $T = \frac{1}{2B}$. Nous supposerons de plus que les fonctions caractéristiques du premier et du second ordre vérifient les hypothèses du paragraphe II.3.2. permettant la simplification des formules (42), (43) et (46). Appliquant les formules simplifiées (44), (45) et (47) aux variables aléatoires que constituent les échantillons du processus X(t), il vient les relations suivantes entre échantillons

de ce processus et ceux du processus X(t) qui s'en déduit par quantification :

$$n, m \in Z \qquad \forall n, \quad E\{\hat{X}(nT)\} = E\{X(nT)\}$$

$$\forall n, \quad E\{\hat{X}^2(nT)\} = E\{X^2(nT)\} + \frac{q^2}{12} \tag{48}$$

$$\forall n, m, m \neq n \quad E\{\hat{X}(nT).\hat{X}(mT)\} = E\{X(nT).X(mT)\}$$

Les deux formules précédentes se condensent d'ailleurs en une seule :

$$\forall n, \forall m \quad E\{\hat{X}(nT).\hat{X}(mT)\} = E\{X(nT).X(mT)\} + \frac{q^2}{12} \cdot \delta_{n-m} \tag{49}$$

Considérons alors un processus B(t) aléatoire, stationnaire, ergodique dont la densité spectrale énergétique est également à support B. borné et dont les moments du premier et deuxième ordre sont donnés par :

$$\forall n, \quad E\{B(nT)\} = 0 \tag{50}$$

$$\forall n,m \quad E\{B(nT).B(mT)\} = \frac{q^2}{12} \cdot \delta_{n-m} \tag{51}$$

Il s'agit donc d'un bruit centré, blanc dans la bande [-B,+B] nous le supposerons de plus indépendant du processus X(t).

Formons le processus $\tilde{X}(t) = X(t) + B(t)$ et calculons ses deux premiers moments en tenant compte des relations (50) et (51)

$$\forall n, \quad E\{\tilde{X}(nT)\} = E\{X(nT)+B(nT)\} = E\{X(nT)\}$$

$$\forall n, \forall m \quad E\{\tilde{X}(nT).\tilde{X}(mT)\} = E\{[X(nT)+B(nT)][X(mT)+B(mT)]\}$$

$$= E\{X(nT).X(mT)\} + \frac{q^2}{12} \cdot \delta_{n-m}$$

Comparant ces formules avec les relations (48) et (49), nous constatons que les processus $\tilde{X}(t)$ et $\hat{X}(t)$ ont même moments du 1er et du 2ème ordre. Remarquons que ce résultat suppose que les hypothèses du paragraphe II.3.2. sont vérifiées c'est-à-dire que le pas de quantification q est suffisamment fin. Nous pouvons donc énoncer la propriété suivante (qui est conforme aux résultats de Gold et Rader, [11]).

Lorsqu'elle est suffisamment fine, la quantification de pas q d'un signal aléatoire à spectre B. borné est statistiquement équivalente au second ordre à l'addition à ce signal d'un bruit centré, blanc dans la bande [-B,+B] de variance $\frac{q^2}{12}$.

Il sera commode de choisir un bruit blanc uniformément réparti entre $-\frac{q}{2}$, $+\frac{q}{2}$ qui répond bien aux conditions précédentes. C'est ce que nous ferons dans le travail de simulation.

III.4 - Erreur due au nombre fini de niveaux

Le calcul précédent suppose que le nombre des niveaux de quantification est

est infini dans le sens positif comme dans le sens négatif. En fait, ce nombre est nécessairement limité à N niveaux, c'est-à-dire que la dynamique d'entrée de quantification sera limitée à ± A = N.q. Toute valeur $|X(t)|>A$ sera quantifiée en $\hat{X} = \pm N.q$, il en résulte une erreur supplémentaire qui a été étudiée et calculée par G. Bonnet [1]. D'après cette étude l'erreur correspondante devient négligeable pour $A \geq 3\sigma$ [12], où σ est l'écart type du signal X(t) à quantifier, c'est-à-dire : $\sigma^2 = E\{(X-m)^2\}$ $(m=E\{X\})$ soit dans le cas où X(t) est aléatoire stationnaire ergodique, ou certain :

$$\sigma^2 = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} [X(t)-m]^2 \, dt$$

III.5 - Remarque

Tous les résultats précédents ont été établis pour la quantification d'arrondi. On pourrait refaire des calculs analogues pour la quantification de seuil, on aboutirait alors à l'équivalence statistique au 2ème ordre avec un bruit uniformément réparti entre -q et 0 ce qui ferait apparaître cette fois l'existence d'un biais de valeur $-\frac{q}{2}$ dans l'estimation de $E\{X\}$ par $E\{\hat{X}\}$.

IV - CALCUL DE L'ERREUR A LA SORTIE D'UN FILTRE NUMERIQUE

IV.1 - Méthode de calcul

D'après l'étude précédente, nous sommes autorisés à remplacer les opérations de quantification, partout où elles ont lieu (à l'entrée du filtre, ou dans la chaîne des calculs pour le cas de l'arrondi sur les multiplications) par des sources de bruit répondant aux conditions fixées ci-dessus.

Nous n'étudierons que le cas où le bruit de simulation est appliqué à l'entrée du filtre ; en effet, on peut toujours se ramener à ce cas en ne considérant que la partie du filtre située en aval du point d'entrée du bruit.

Le traitement réel correspond donc au schéma de la figure 4.5 :

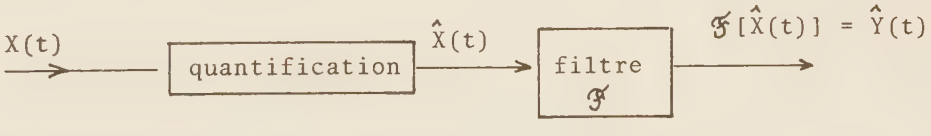


Figure 4.5

Nous le remplacerons par le schéma de la figure 4.6 qui est équivalent statistiquement (au 1er et 2ème ordre) :

Figure 4.6

---

12) Pour une variable aléatoire Gaussienne l'erreur relative due à la limitation de la dynamique du quantificateur à $A = 3\sigma$ est de l'ordre de 10^{-3}

L'erreur réelle en sortie du filtre sera : $\hat{\varepsilon}(t) = \hat{Y}(t) - Y(t)$, $\left(Y(t) = \mathcal{F}[X(t)]\,,\ \hat{Y}(t) = \mathcal{F}[\hat{X}(t)]\right)$. Nous caractériserons cette erreur par ses moments du 1er et 2ème ordre : $b_S = E\{\varepsilon(t)\}$ appelé biais et $\sigma_S^2 = E\{[\hat{\varepsilon}(t) - b_S]^2\}$ qui est la variance de l'erreur.

σ_S caractérise l'importance de l'erreur, tandis que b_S précise la manière dont dont elle se répartit en moyenne autour de la valeur exacte.

Pour calculer σ_S et b_S nous utiliserons les propriétés d'équivalence statistique du schéma équivalent, ceci nous permet d'écrire :

$$\begin{aligned} b_S &= E\{\tilde{\varepsilon}(t)\} \\ \sigma_S^2 &= E\left\{[\tilde{\varepsilon}(t) - b_S]^2\right\} \end{aligned} \qquad \text{avec} \quad \tilde{\varepsilon}(t) = \tilde{Y}(t) - Y(t)$$

Le problème devient alors linéaire, car on a :

$$\tilde{\varepsilon}(t) = \mathcal{F}[\tilde{X}(t)] - \mathcal{F}[X(t)] = \mathcal{F}[X(t)+B(t)] - \mathcal{F}[X(t)] = \mathcal{F}[B(t)]$$

puisque $\mathcal{F}$ est linéaire. Il vient donc finalement :

$$b_S = E\{\mathcal{F}[B(t)]\} \tag{52}$$

$$\sigma_S^2 = E\left\{\left|\mathcal{F}[B(t)] - b_S\right|^2\right\} \tag{53}$$

IV.2 - Expression de l'erreur en sortie

Nous prendrons pour B(t) un bruit blanc dans la bande [-B,+B] uniformément réparti entre $-\frac{q}{2}$ et $+\frac{q}{2}$ qui satisfait bien aux hypothèses que nous avons déjà vu ; soit H(t) la réponse impulsionnelle du filtre $\mathcal{F}$, $\Gamma_{BE}(\tau)$ la fonction de corrélation de B(t) et $\Gamma_{BS}(\tau)$ celle de $\mathcal{F}[B(t)]$. D'après les formules des interférences nous avons :

- au 1er ordre : $E\{\mathcal{F}[B(t)]\} = \mathcal{F}[E\{B(t)\}] = \mathcal{F}[o] = 0$

donc : $b_S = 0$ (54)

- au 2ème ordre : $\Gamma_{BS}(\tau) = \Gamma_{BE} * \Gamma_H \quad (\Gamma_H(\tau) = H * H^{\#},\ H^{\#}(t) = H(-t))$

Or $\Gamma_{BE}(\tau) = E\{B^2(t)\}.\text{sinc } 2\pi B\tau = \frac{q^2}{12}.\text{sinc } 2\pi B\tau$

et $H(t) \in BL^2 \rightarrow \Gamma_H(\tau) \in BL^2$ (voir § 1.I.2.3)

enfin on sait que sinc $2\pi B\tau$ est l'élément neutre de la convolution dans BL^2, donc :

$$\Gamma_{BS}(\tau) = \frac{q^2}{12}\,\Gamma_H(\tau)$$

pour $\tau = 0$ $\qquad \Gamma_{BS}(0) = \sigma_S^2$

$$\Gamma_H(o) = (H*H^{\#})_{\tau=o} = \int_{-\infty}^{+\infty} |H(t)|^2.dt = \int_{-\infty}^{+\infty} |h(\nu)|^2 \, d\nu = ||H||^2$$

d'où finalement : $$\sigma_S^2 = \frac{q^2}{12} \cdot ||H||^2 \qquad (55)$$

ou : $$\sigma_S = \frac{q}{2\sqrt{3}} \cdot ||H|| \qquad (55')$$

Les formules (54) et (55') nous permettent d'énoncer :

Une quantification d'arrondi de pas q à l'entrée d'un filtre n'introduit pas de biais en sortie mais une erreur dont l'écart type (ou dispersion autour de la valeur exacte) est proportionnel au pas de quantification q est à la norme du filtre [13].

IV.3 - Application aux filtres numériques récursifs du 2ème ordre

L'étude du filtre du 2ème ordre est très importante car, comme nous l'avons vu, il s'agit de l'élément de base des structures en cascade ou en parallèle. Nous allons calculer $||H||$ en fonction des paramètres du filtre, ce qui nous permettra de connaître leur influence sur le comportement du filtre tant pour le bruit de quantification que pour la stabilité.

Soit donc un filtre dont la fonction de transfert en z s'écrit :

$$H(z) = \frac{A + B.z^{-1}}{1 + C.z^{-1} + D.z^{-2}} \qquad (56)$$

Nous utiliserons la formule (25) du paragraphe 2.IV.3.4. (en prenant pour unité la largeur de bande 2B), et nous calculerons l'intégrale qui y figure par la méthode des résidus. Les pôles de H(z) se trouvant à l'intérieur du cercle unité, ceux de $H(\frac{1}{z})$ se trouvent à l'extérieur,

on a donc : $$||H||^2 = R_o + R_1 + R_2$$

R_o et R_1 étant les résidus de $z^{-1}.H(z).H(\frac{1}{z})$ relatifs aux pôles z_o et z_1 de H(z) et R_2 le résidu relatif au pôle z = o.

Nous avons : $$H(z) = \frac{A.z^2+B.z}{(z-z_o)(z-z_1)} \qquad H(\frac{1}{z}) = \frac{A + B.z}{z_o z_1 \left(z-\frac{1}{z_o}\right)\left(z-\frac{1}{z_1}\right)}$$

---

13) Dans le cas d'une quantification de seuil, la seule différence serait un biais en sortie de valeur $b_S = -\frac{q}{2}$ (d'après § II.5).

$$z^{-1}.H(z).H\left(\frac{1}{z}\right) = \frac{(Az+B)\ \ (A+Bz)}{z_o\ z_1\ (z-z_o)\ \ (z-z_1)\ \ \left(z-\frac{1}{z_o}\right)\ \ \left(z-\frac{1}{z_1}\right)}$$

$$R_o = \frac{(Az_o+B)\ \ (A+Bz_o)}{z_o\ z_1\ (z_o-z_1)\left(z_o-\frac{1}{z_o}\right)\ \ \left(z_o-\frac{1}{z_1}\right)}$$

$$R_1 = \frac{(Az_1+B)\ \ (A+Bz_1)}{z_o\ z_1\ (z_1-z_o)\ \ \left(z_1-\frac{1}{z_o}\right)\left(z_1-\frac{1}{z_1}\right)}$$

Nous constatons que R_o+R_1 est une fonction symétrique des racines. On pourra donc l'exprimer en fonction de A, B, C et D.

$$R_o+R_1 = \frac{(Az_o+B)(Bz_o+A)(z_1-z_o)\left(z_1-\frac{1}{z_1}\right)\left(z_1-\frac{1}{z_o}\right)+\ (Az_1+B)(Bz_1+A)(z_o-z_1)\left(z_o-\frac{1}{z_o}\right)\left(z_o-\frac{1}{z_1}\right)}{z_oz_1(z_o-z_1)\ \ \left(z_o-\frac{1}{z_o}\right)\left(z_1-\frac{1}{z_1}\right)\left(z_o-\frac{1}{z_1}\right)\left(z_1-\frac{1}{z_o}\right)} = \frac{\eta}{\delta}$$

$$\delta = -\ (z_oz_1-1)^2(z_o-z_1)^2\left(z_oz_1+\frac{1}{z_oz_1}-\frac{z_1^2+z_o^2}{z_o\ z_1}\right) = -\ \frac{(C^2-4D)(1-D)^2\ \left[(1+D)^2-C^2\right]}{D}$$

$$\eta = -\ \frac{(z_oz_1-1)(z_o-z_1)}{z_o\ z_1}\ \Big[AB(z_o^2+1)(z_1^2-1)+(A^2+B^2)z_o(z_1^2-1)-AB(z_1^2+1)(z_o^2+1)$$
$$-(A^2+B^2)z_1(z_o^2-1)\Big]$$

$$\eta = \frac{(z_oz_1-1)(z_o-z_1)^2}{z_o\ z_1}\ \left[2\ AB(z_o+z_1)\ +\ (A^2+B^2)\ \ (z_oz_1+1)\right]$$

$$\eta = -\ \frac{(C^2-4D)\ \ (1-D)}{D}\left[(1+D)\ \ (A^2+B^2)-2\ ABC\right]$$

d'où : $$\frac{\eta}{\delta} = \frac{(1+D)\ \ (A^2+B^2)\ -\ 2\ ABC}{(1-D)^2\ \ (1+D+C)\ \ (1+D-C)}$$

soit finalement : $$||H||^2 = \frac{(1+D)\ \ (A^2+B^2)-2\ A.B.C}{(1-D)^2\ \ (1+D+C)\,(1+D-C)} \qquad (57)$$

V - ETUDE DE LA STABILITE D'UN FILTRE NUMERIQUE RECURSIF

V.1 - Généralités

Nous avons vu au premier paragraphe de ce chapitre que certaines erreurs

étaient liées à la stabilité. Après avoir calculé ces erreurs nous allons étudier la stabilité, ainsi que la précision nécessaire sur les coefficients pour éviter des oscillations spontanées, principal obstacle au bon fonctionnement des filtres numériques récursifs.

V.2 - Stabilité et sensibilité au bruit de quantification

Comme nous l'avons vu au paragraphe I.4.3. l'un des facteurs déterminant le degré de stabilité d'un filtre est la distance des pôles au cercle unité. La fonction de transfert en z d'un filtre numérique est une fraction rationnelle en z qui peut s'écrire : $H(z) = \frac{P(z)}{Q(z)}$ avec $Q(z) = (z-z_o)(z-z_1)\ldots(z-z_{N-1})$ où $z_o, z_1, \ldots, z_{N-1}$ sont les N pôles du filtre. Nous avons vu (§ IV.2.1.) que le gain complexe est alors donné par : $h(\nu) = H(z)_{z=e^{2\pi i\frac{\nu}{2B}}}$, si donc l'un des pôles est proche du cercle unité, $h(\nu)$ présentera à la fréquence correspondante un maximum très aigu. La quantité $\sup_{\nu\in[-B,+B]} |h(\nu)|$ peut donc nous renseigner sur le degré de stabilité du filtre.

Ainsi le degré de stabilité d'un filtre qui est une propriété sélective (c'est-à-dire qu'elle dépend de la fréquence) ne répond pas au même critère que la sensibilité au bruit de quantification qui en tant que bruit blanc peut s'évaluer à l'aide d'un paramètre global (c'est-à-dire faisant intervenir le comportement du filtre sur tout l'axe des fréquences) tel que $||H||$.

Nous formulerons ici deux remarques :

1° - Le paramètre $\sup_{\nu\in[-B,+B]} |h(\nu)|$ dépend également d'un facteur d'échelle qui caractérise l'amplification du filtre (on peut prendre par exemple le gain en continu $|h(o)|$) et qui n'a rien à voir avec le degré de stabilité. Ce paramètre ne fournit donc qu'une indication relative sur la stabilité du filtre.

2° - Si $\sup_{\nu\in[-B,+B]} |h(\nu)|$ est grand, il est assez probable que $||H||^2 = \int_{-\infty}^{+\infty} |h(\nu)|^2 \, d\nu$ le soit aussi. Le degré de stabilité n'est donc pas totalement indépendant de la sensibilité au bruit de quantification

En résumé, nous pourrons dire que le degré de stabilité décroît quand la densité maximale d'énergie correspondant aux fréquences propres (ou fréquences de résonnance) du filtre croît, tandis que l'importance de l'erreur en sortie croît avec l'énergie totale de la réponse impulsionnelle.

V.3 - Domaine de stabilité en fonction des coefficients

Un filtre numérique récursif d'ordre N possède une fonction de transfert en z dont le dénominateur est un polynôme en z de degré N :

$\mathcal{D}(z) = \sum_{n=o}^{N} b_i . z^{N-i}$. Les racines de ce polynôme, qui constituent les pôles du filtre, sont des fonctions continues de ses N coefficients $b_1, b_2, \ldots, b_N$ ($b_o=1$). La condition de stabilité imposant que le module de ces N racines soit inférieur à l'unité elle se traduit par N relations d'inégalité entre les coefficients, ce

qui revient à définir un domaine de l'espace N - dimensionnel constitué par l'ensemble $(b_1, b_2, \ldots, b_N)$. Tout point de ce domaine correspond de par sa définition à un filtre stable d'ordre N. C'est pourquoi nous l'appellerons domaine de stabilité du filtre d'ordre N.

On conçoit que le degré de stabilité de ce filtre sera d'autant plus élevé que le point qui lui correspond dans le domaine de stabilité sera plus éloigné des frontières de ce domaine, c'est-à-dire qu'une faible variation de l'un quelconque des coefficients ne suffira pas à l'en faire sortir.

Dans le cas où le filtre est décomposé en éléments du second ordre associés en parallèle ou en cascade, la condition nécessaire et suffisante de stabilité est que chacun des filtres élémentaires soit stable. Le domaine de stabilité se réduit donc à la réunion des domaines de stabilités bidimensionnels correspondant à tous les éléments du second ordre. Il en résulte que l'on pourra connaître la stabilité de l'ensemble en étudiant séparément la stabilité de chacun des éléments.

Ce cas étant le plus fréquent (la structure décomposée en éléments du 2ème ordre est d'ailleurs la seule présentant un degré de stabilité suffisant dans le cas des filtres d'ordre élevé) nous allons étudier en détail le domaine de stabilité des filtres du 2ème ordre.

V.4 - Domaine de stabilité des filtres du 2ème ordre

$$\text{On a :} \quad H(z) = \frac{A + Bz^{-1}}{1+Cz^{-1}+Dz^{-2}} = \frac{Az^2+Bz}{z^2+Cz+D}$$

$$\mathcal{D}(z) = z^2 + Cz + D$$

1° - Cas des pôles imaginaires : $C^2-4D < 0$

On a : $z_1=\bar{z}_o \rightarrow |z_1|=|z_o|$ et $|z_o|^2 = z_o.\bar{z}_o = z_o z_1 = D$

Par suite : $(|z_o|<1) \text{ n } (|z_1|<1) \leftrightarrow |z_o|<1 \leftrightarrow |z_o|^2<1 \leftrightarrow D<1$ (58)

2° - Cas des pôles réels : $C^2-4D \geqslant 0$

$$z_o = \frac{1}{2}\left[-C+\sqrt{C^2-4D}\right]$$

$$z_1 = \frac{1}{2}\left[-C-\sqrt{C^2-4D}\right]$$

$$|z_o|<1 \leftrightarrow (z_o<1) \text{ n } (z_o>-1)$$

a) $z_o < 1 \leftrightarrow \frac{1}{2}\left[-C+\sqrt{C^2-4D}\right]< 1$

$\leftrightarrow \sqrt{C^2-4D} < C+2$

$\leftrightarrow (C+2>0) \text{ n } (C^2-4D < C^2+4C+4)$

$\leftrightarrow (C>-2) \text{ n } (C+D > -1)$

FILTRE RECURSIF
DU 2è ORDRE
DOMAINE DE STABILITE

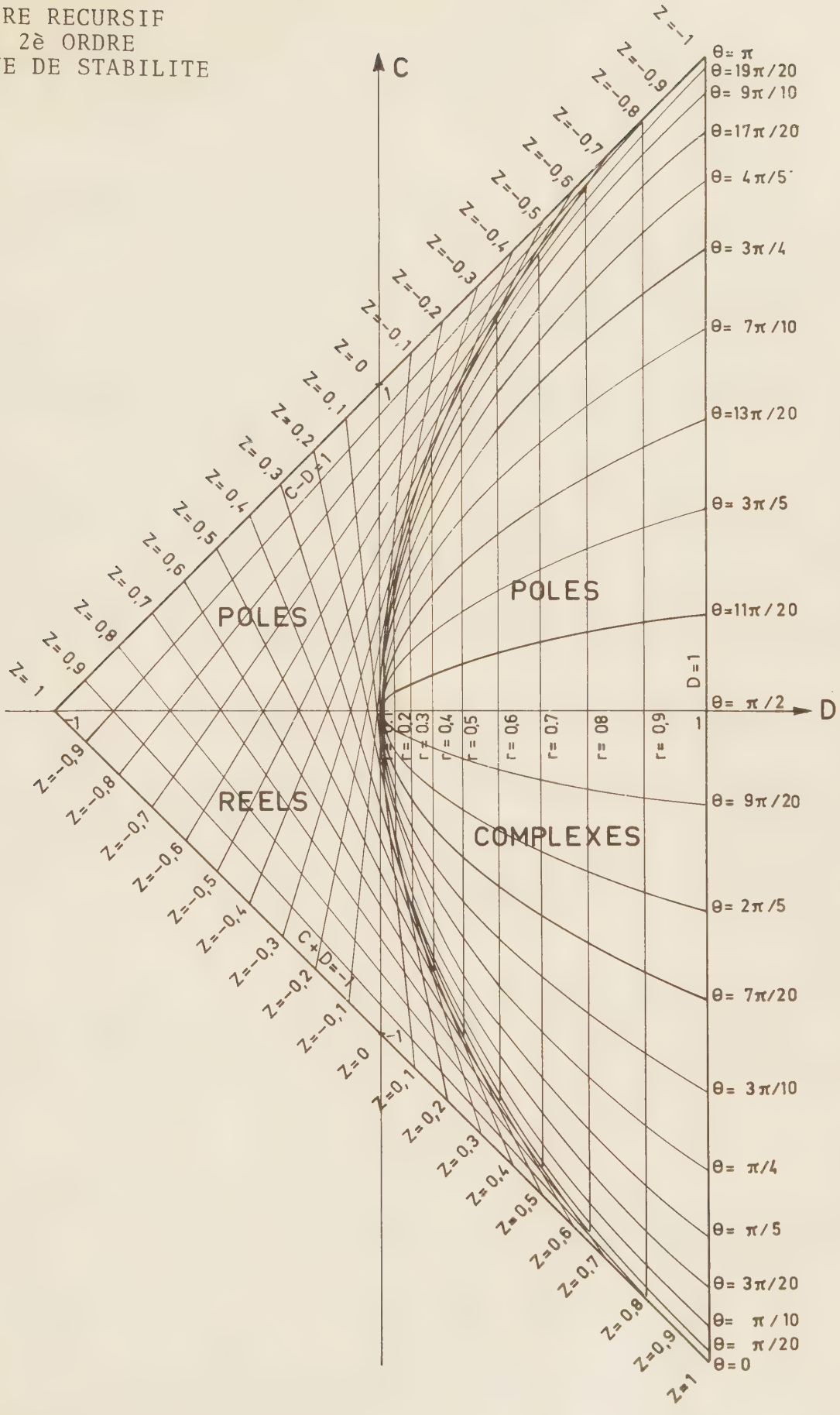


Figure 4.7

b) $z_o > -1 \leftrightarrow \frac{1}{2}\left[-C+\sqrt{C^2-4D}\right] > -1$

$\leftrightarrow \sqrt{C^2-4D} > C-2$

$\leftrightarrow$ (C<2) U (C-D>1)

Soit finalement : $|z_o|<1 \leftrightarrow$ (C>-2) n (C+D>-1) n [(C-D>-1) U (C<2)]

de même on démontre : $|z_1|<1 \leftrightarrow$ (C<2) n (C-D<1) n [(C+D<-1) U (C>-2)]

d'où : $(|z_o|<1$ n $|z_1|<1) \leftrightarrow$ (-2<C<2) n (C+D>-1) n (C-D<1) (59)

Les expressions (58) et (59) nous permettent de construire le graphe de la figure 4.7 qui donne le domaine de stabilité dans le plan (C,D) partagé en deux sous domaines correspondant aux pôles réels et aux pôles imaginaires.

Sur le même graphe, on a également tracé les courbes z_o = Cte, z_1 = Cte dans le cas des pôles réels et r = Cte, θ = Cte dans le cas des pôles complexes $z = re^{\pm i\theta}$. Ceci permet, connaissant les deux pôles d'en déduire graphiquement les coefficients C et D et le point représentatif correspondant dans le domaine de stabilité, qui, de par sa distance aux frontières de ce domaine nous permet d'évaluer le degré de stabilité du filtre (en particulier on pourra mesurer les erreurs maximales autorisées sur C et D pour ne pas sortie du domaine de stabilité).

VI - ASSOCIATION DE FILTRES EN PARALLELES OU EN CASCADE

Nous avons vu au paragraphe 4.V.3. que pour des raisons de stabilité, on avait intérêt à décomposer les filtres d'ordre élevé en filtres du 2ème ordre associés en parallèle ou en cascade. Nous allons étudier ces associations du point de vue des erreurs en sortie, ce qui, d'après la relation (55') revient à étudier le module de l'ensemble en fonction des modules de chacun des filtres ainsi associés.

VI.1 - Association en parallèle

Soient $H_1(t)$, $H_2(t)$,..., $H_N(t)$ les réponses impulsionnelles de N filtres associés en parallèle. La réponse impulsionnelle H(t) de l'ensemble s'écrit :

$$H(t) = \sum_{i=1}^{N} H_i(t)$$

Or, H(t) et les $H_i(t)$ sont des éléments de l'espace vectoriel normé BL^2 on peut donc écrire d'après l'inégalité des Minkoswski :

$$||H|| \leqslant \sum_{i=1}^{N} ||H_i|| \qquad (60)$$

On peut donc, compte-tenu de la relation (55') énoncer :

L'erreur (due à la quantification du signal d'entrée) à la sortie de N filtres associés en parallèle est au plus égale à la somme des erreurs en sortie qui existerait pour chacun des filtres pris isolément. L'égalité stricte à lieu lorsque les filtres sont identiques à un coefficient multiplicatif près, l'erreur en sortie est alors proportionnelle à N.

VI.2 - Association en cascade

Soient H_1 et H_2 deux filtres associés en cascade, on sait que la réponse impulsionnelle H de l'ensemble s'écrit : $H = H_1 * H_2$. Or, l'opération de convolution n'est pas une opération propre de la structure d'espace vectoriel et il n'existe pas de relation analogue à la relation (60). La valeur de $\|H\|$ par rapport à $\|H_1\|$ et $\|H_2\|$ dépend de $H_1(t)$ et $H_2(t)$ et donc le rapport entre l'erreur en sortie de H les erreurs en sortie de H_1 et de H_2 seuls est une affaire de cas d'espèce : il n'existe pas de règle générale.

Chapitre 5

PROJET DE REALISATION D'UN FILTRE NUMERIQUE RECURSIF PROGRAMMABLE

I - DESCRIPTION ET INTERET DE L'APPAREIL PROJETE

L'appareil que nous proposons est un filtre dont on pourra modeler à volonté la courbe de réponse par simple affichage des valeurs des coefficients qui la caractérisent. Le choix d'un appareil fonctionnant selon le principe des filtres numériques récursifs nous a paru particulièrement bien adapté à cette fonction de "filtre programmable". En effet, le petit nombre d'opérations arithmétiques répétitives intervenant dans ce type de filtre convient particulièrement bien à une réalisation câblée du type d'un petit calculateur spécialisé et le nombre relativement faible des coefficients caractérisant la réponse permet un affichage rapide et commode de celle-ci.

Avant d'étudier le détail de la réalisation technologique, il nous faudra aborder deux points importants : à savoir, le choix de l'algorithme correspondant à une structure de filtre numérique récursif qui soit aussi stable que possible, et la précision nécessaire de l'unité arithmétique compte-tenu de l'erreur tolérée en sortie et des risques d'instabilité engendrés par les imprécisions sur les coefficients ou dans les calculs.

II - CHOIX DE LA STRUCTURE

Bien entendu, nous utiliserons une structure décomposée en éléments du second ordre, nous étudierons donc successivement, le mode d'association choisi et la structure de ces éléments.

II.1 - Choix du mode d'association

Le filtre projeté ayant des coefficients que l'on pourra faire varier à volonté, nous préférerons l'association en parallèle pour laquelle la seule donnée des modules des éléments du 2ème ordre nous permet de calculer un majorant de l'erreur en sortie du filtre global. L'association en série où cette erreur dépend de tous les coefficients de chacun des éléments, nécessiterait pour chaque ensemble de coefficients choisi un calcul permettant de s'assurer que l'erreur en sortie n'est pas trop importante.

II.2 - Structure des éléments du 2ème ordre

Les deux structures possibles pour les éléments du 2ème ordre sont la structure directe et la structure canonique. Nous choisirons celle qui provoque l'erreur en sortie la plus faible. Cette erreur en sortie aura pour causes :

1° - L'erreur due à la quantification de pas q du signal d'entrée ;

2° - Les erreurs d'arrondi après chaque multiplication. En effet, le produit de deux valeurs quantifiées à N bits donne un résultat à 2N bits que l'on ramène à N bits en abandonnant les N bits les moins significatifs, cette opération équivaut donc à une nouvelle quantification de pas q.

Conformément à l'étude précédente, nous remplacerons chaque quantification de pas q par une source de bruit blanc de variance $\frac{q^2}{12}$, les résultats en sortie différeront en raison de la position différente de ces sources de bruits selon les structures.

Les figures 5.1 et 5.2 donnent la répartition des sources de bruit respectivement dans les structures directe et canonique. Ces sources de bruit peuvent se regrouper en quelques points de ces structures ainsi que le montrent les figures 5.3. et 5.4. La fonction de transfert correspondant à la fraction de filtre placée entre ces sources et la sortie est indiquée sous chaque figure, enfin, les modules de ces fonctions de transfert se déduisent tous de la relation (57) du paragraphe 4.IV.3.

On a donc, avec les notations des figures les variances d'erreur en sortie suivantes :

- pour la forme directe :

$$\sigma^2_{SD} = \frac{q^2}{12} \left| ||H||^2 + 4||H_o||^2 \right| \qquad (61)$$

- pour la forme canonique :

$$\sigma^2_{SC} = \frac{q^2}{12} \left| ||H||^2 + \frac{3}{A^2} ||H||^2 + 1 \right| \qquad (62)$$

Etudions le signe de la quantité :

$$\varepsilon = \frac{1}{\left(\frac{q^2}{12}\right)} \left[\sigma^2_{SC} - \sigma^2_{SD} \right] = \frac{3}{A^2} ||H||^2 + 1 - 4 ||H_o||^2$$

soit, compte tenu des expressions de $||H||$ et de $||H_o||$:

$$\varepsilon = \frac{1}{\Delta} \left| 3(1+D) \left(1+\frac{B^2}{A^2}\right) - \frac{6BC}{A} - 4(1+D) \right| + 1$$

$$\varepsilon = \frac{1}{\Delta} \left| (1+D) \left(3\frac{B^2}{A^2} - 1\right) - 6 \frac{B}{A}.C \right| + 1$$

où l'on a posé : $\Delta = (1-D)^2 (1+D+C) (1+D-C)$ dénominateur de $||H||^2$ (remarquons que pour un filtre stable, on a $\Delta > 0$).

Dans le cas où les pôles sont assez proches du cercle unité (cas le plus

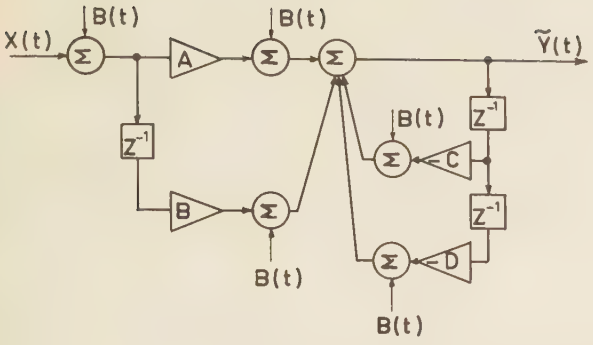


Figure 5.1

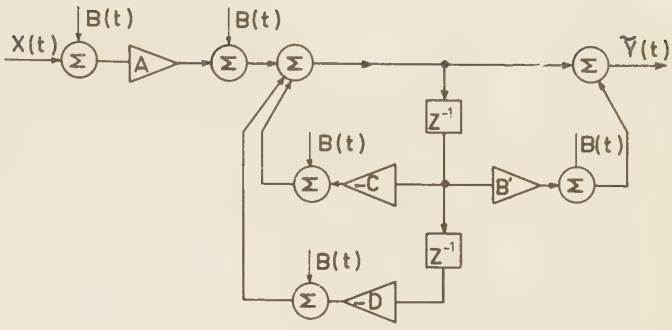


Figure 5.2

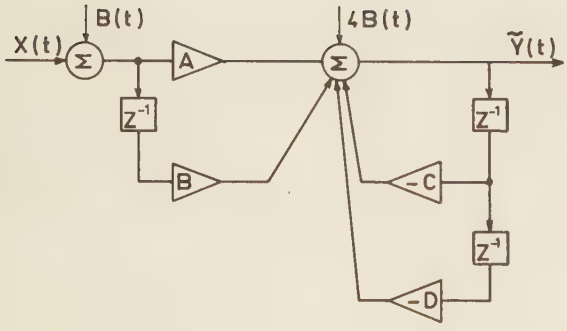


Figure 5.3

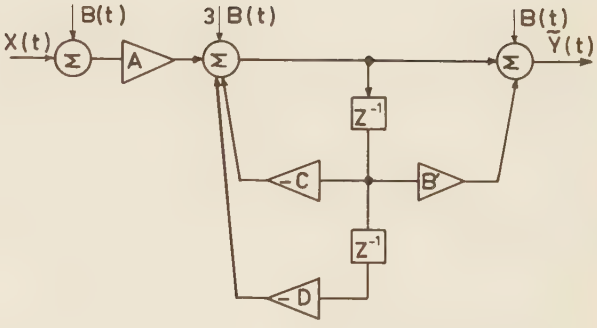
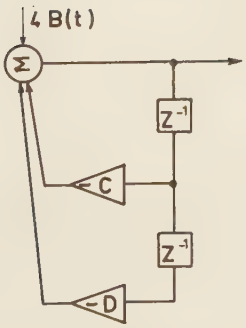


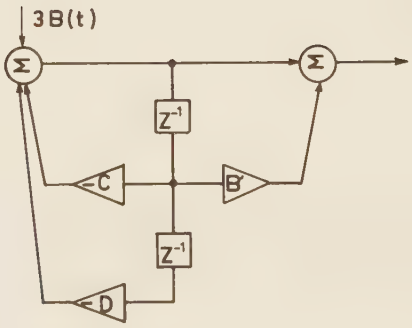
Figure 5.4



$$H_o(z) = \frac{1}{1+Cz^{-1}+Dz^{-2}}$$

$$||H_o||^2 = \frac{1+D}{(1-D)^2(1+D+C)(1+D-C)}$$

Figure 5.5



$$H_1(z) = \frac{1}{A} H(z)$$

$$||H_1||^2 = \frac{(1+D)(1+B'^2)-2B'C}{(1-D)^2(1+D+C)(1+D-C)}$$

Figure 5.6

fréquent dans les filtres performents, et aussi le plus délicat) le point représentatif dans le domaine de stabilité est proche des frontières de ce domaine et Δ est petit devant l'unité ; le signe de ε est alors celui de la quantité entre crochets, il dépend donc essentiellement de la valeur des coefficients. Ceux-ci étant variables dans l'appareil projeté, il est impossible de choisir une structure qui dans tous les cas minimise l'erreur en sortie. Il semble cependant que dans la majorité des filtres courants nous aurons $\varepsilon < 0$, nous choisirons donc finalement la structure canonique.

III - ERREUR EN SORTIE ET STABILITE AVEC LA STRUCTURE CHOISIE

III.1 - Erreur en sortie

Nous essaierons d'évaluer une borne supérieure de l'erreur relative en sortie en fonction du nombre P de bits de l'unité arithmétique, et de l'ordre N du filtre.

La valeur de l'erreur absolue en sortie pour un élément du 2ème ordre, compte tenu de la structure choisie est donnée par la formule (62)

Soit σ_X l'écart type du signal d'entrée additionné du bruit équivalent à la quantification. Pour réduire suffisamment l'erreur due au nombre fini de niveaux de quantification nous supposerons $\sigma_X \simeq \alpha/3$ où α est la dynamique d'entrée (amplitude maximale pouvant être quantifiée sans écrétage) [14].

Nous prendrons pour signal type un bruit blanc dans la bande [-B,+B] dans ces conditions nous avons vu que la variance du signal de sortie correspondant s'écrit :

$$\sigma_Y = ||H||.\sigma_X \qquad (63)$$

L'erreur relative en sortie a pour expression :

$$\varepsilon_r = \frac{\sigma_{SC}}{\sigma_Y}$$

d'où, compte-tenu de la relation $q = \frac{\alpha}{2^P}$ (P nombre de bits de l'unité arithmétique, bit de signe non compris) et des relations (62) et (63) :

$$\varepsilon_r = \frac{\sqrt{3}}{2^{P+1}} \sqrt{1+\frac{3}{A^2}+\frac{1}{||H||^2}}$$

1° - Nous aurons en général $||H|| \geqslant 1$ et l'influence de l'erreur d'arrondi dans la multiplication par B' sera peu importante par rapport à celle des erreurs d'arrondi dans les multiplications par A, C et D.

2° - Nous choisirons aussi $A \geqslant 1$ pour que l'influence de l'erreur de quantification du signal d'entrée soit du même ordre que celle des erreurs d'arrondi dans les multiplications par A, C et D.

---

14) Dans ces conditions on montre [1] que l'erreur d'écrétage ne dépasse pas 10^{-3}.

Dans ces conditions, nous aurons :

$$\varepsilon_r \leqslant \frac{\sqrt{15}}{2^{P+1}} \simeq \frac{1}{2^{P-1}}$$

Soit, finalement pour un filtre d'ordre N associant $\frac{N}{2}$ cellules du 2ème ordre en parallèle :

$$\varepsilon_{rN} \leqslant \frac{N}{2^P} \qquad (64)$$

III.2 - Stabilité

Ainsi que nous l'avons vu au paragraphe 4.V.3 la condition nécessaire et suffisante de stabilité d'un filtre d'ordre N est que chacun de ses éléments du 2ème ordre soit stable. Nous avons vu également que la condition de stabilité se traduit par l'appartenance des coefficients C et D de chaque élément du 2ème ordre à un domaine du plan C.D. (voir figure 4.7). En fait ces coefficients étant quantifiés, ils sont entachés d'une erreur absolue positive ou négative au plus égale à $\frac{q}{2}$ (dans le cas de la quantification d'arrondi), on réduira donc le domaine de stabilité du plan C.D. en tenant compte de cette erreur lorsqu'elle joue dans le sens défavorable. Nous savons que le domaine de stabilité d'un élément du 2ème ordre est défini par :

$$\begin{aligned} C - D &\leqslant 1 \\ C + D &\geqslant -1 \\ D &\leqslant 1 \end{aligned}$$

Si C_o et D_o sont les valeurs théoriques de C et D respectivement quantifiées avec les pas q et q', on aura : $C_o - \frac{q}{2} \leqslant C \leqslant C_o + \frac{q}{2}$ et $D_o - \frac{q'}{2} \leqslant D \leqslant D_o + \frac{q'}{2}$, d'où le domaine de stabilité sûre pour les valeurs théoriques C_o et D_o (l'erreur de quantification intervenant dans le sens le plus défavorable).

$$D_o \leqslant 1-\frac{q'}{2}$$

$$C_o - D_o \leqslant 1-\frac{q+q'}{2}$$

$$C_o + D_o \geqslant -1+\frac{q+q'}{2}$$

Dans le cas d'une arithmétique à P bits (bit de signe non compris), l'amplitude maximale en valeur absolue de D étant égale à 1 et celle de C à 2, on a : $q = \frac{2}{2^P}$ et $q' = \frac{1}{2^P}$ d'où finalement le domaine de stabilité réel :

$$D_o \leqslant 1-\frac{1}{2^{P-1}}$$

$$C_o - D_o \leqslant 1-\frac{3}{2^P}$$

$$C_o + D_o \geqslant -1+\frac{3}{2^P}$$

Ce domaine se réduit donc à un triangle intérieur au triangle de stabilité théorique de la figure 4.7.

IV - STRUCTURE TECHNOLOGIQUE DE L'APPAREIL

IV.1 - Algorithme de calcul

La structure finalement retenue est celle de la figure 5.7 et dont la relation entrée-sortie sous forme de transformée en Z se caractérise par les relations suivantes :

$$T_i(z) = A_i.X(z) - C_i.z^{-1}.T_i(z) - D_i.z^{-2}.T_i(z)$$

$$Y_i(z) = T_i(z) + B'_i.z^{-1}.T_i(z) \qquad (63)$$

$$Y(z) = \sum_{i=1}^{P} Y_i(z)$$

qui correspondent aux fonctions de transfert :

$$H_i(z) = \frac{A_i + B_i.z^{-1}}{1-C_i.z^{-1}-D_i.z^{-2}} = A.\ \frac{1 + B'_i.z^{-1}}{1-C_i.z^{-1}-D_i.z^{-2}}$$

de filtres associés en parallèle, dont la fonction de transfert globale en z est :

$$H(z) = \sum_{i=1}^{P} H_i(z)$$

Les relations (63) peuvent encore s'écrire sous la forme suivante :

$$T_i(z) = A_i.X(z) + C'_i.z^{-1}.T_i(z) + D'_i.z^{-2}.T_i(z)$$

$$T(z) = \sum_{i=1}^{P} (T_i(z) + B'_i.z^{-1}.T_i(z)) \qquad (63')$$

avec : $C'_i = -C_i$; $D'_i = -D_i$; $B'_i = B_i/A_i$.

Cet algorithme fait intervenir les éléments suivants :

1° - 2P registres permettant la prise en mémoire de T_i sur deux cycles de calcul :
T_{1i}(i=1àP) correspond à $z^{-1}.T_i(z)$ dans les équations (63')
T_{2i}(i=1àP) correspond à $z^{-2}.T_i(z)$ dans les équations (63').

2° - Deux registres R_1 et R_2 présentant au multiplieur les deux facteurs de chaque produit à effectuer.

3° - Deux registres accumulateurs R'_a et R_a situés respectivement à la sortie du multiplieur et à la sortie de l'ensemble.

L'algorithme utilisé est alors indiqué dans le tableau 5.1, les indications des colonnes de droite sont expliquées au paragraphe suivant.

IV.2 - Schéma synoptique de l'appareil (figure 5.8)

Ainsi que le montre le schéma synoptique de la figure 5.8 l'appareil est organisé autour de deux mémoires circulantes, l'une M_1 à accès exclusivement séquentiel est la mémoire correspondant aux registres T_{1i} et T_{2i} (i=1àP), l'autre M_2 à accès séquentiel ou aléatoire contenant les coefficients A_i, B'_i, C'_i, D'_i (i=1àP).

L'ordre de classement des coefficients dans la mémoire M_2, l'organisation de la mémoire M_1 (comprenant un registre séparé et les inverseurs électroniques I et II) ainsi qu'un aiguillage électronique III constituent une organisation fonctionnant automatiquement selon l'algorithme du tableau 5.1. On pourra le vérifier à partir du contenu des mémoires M_1 et M_2 en début de séquence (figure 5.9) et de la position des inverseurs électroniques I, II et III au cours de la séquence donnée en regard de l'algorithme sur le tableau 5.1.

On remarquera en particulier qu'une séquence se compose de sous-séquences correspondant chacune au calcul de l'un des filtres du 2ème ordre associés en parallèle. Chaque sous-séquence comprend 4 coups d'horloge (notés 0, 1, 2, 3) correspondant à l'ensemble des opérations de la boucle sur i (i=1àP) de l'algorithme.

On notera qu'en fin de séquence un coup d'horloge supplémentaire (notée 00) permet la remise en phase des mémoires M_1 et M_2 avant d'attaquer la séquence suivante.

On appelle séquence, l'ensemble des opérations permettant le calcul d'un échantillon de sortie. L'appareil travaillant en temps réel, la durée totale d'une séquence est inférieure à la période d'échantillonnage.

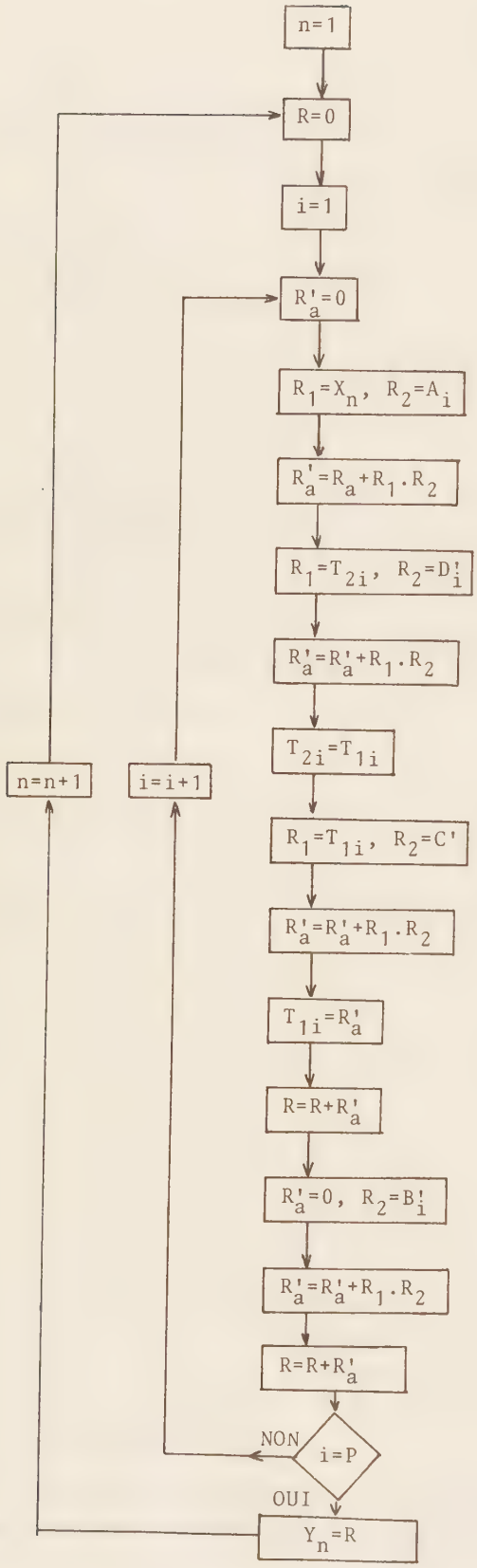


Tableau 5.1

I	II	III	Impulsions d'horloge
1	3	$\overline{0}$	1
1	4	$\overline{0}$	2
2	4	$\overline{0}$	3
1	4	0	0
1	0	0	00

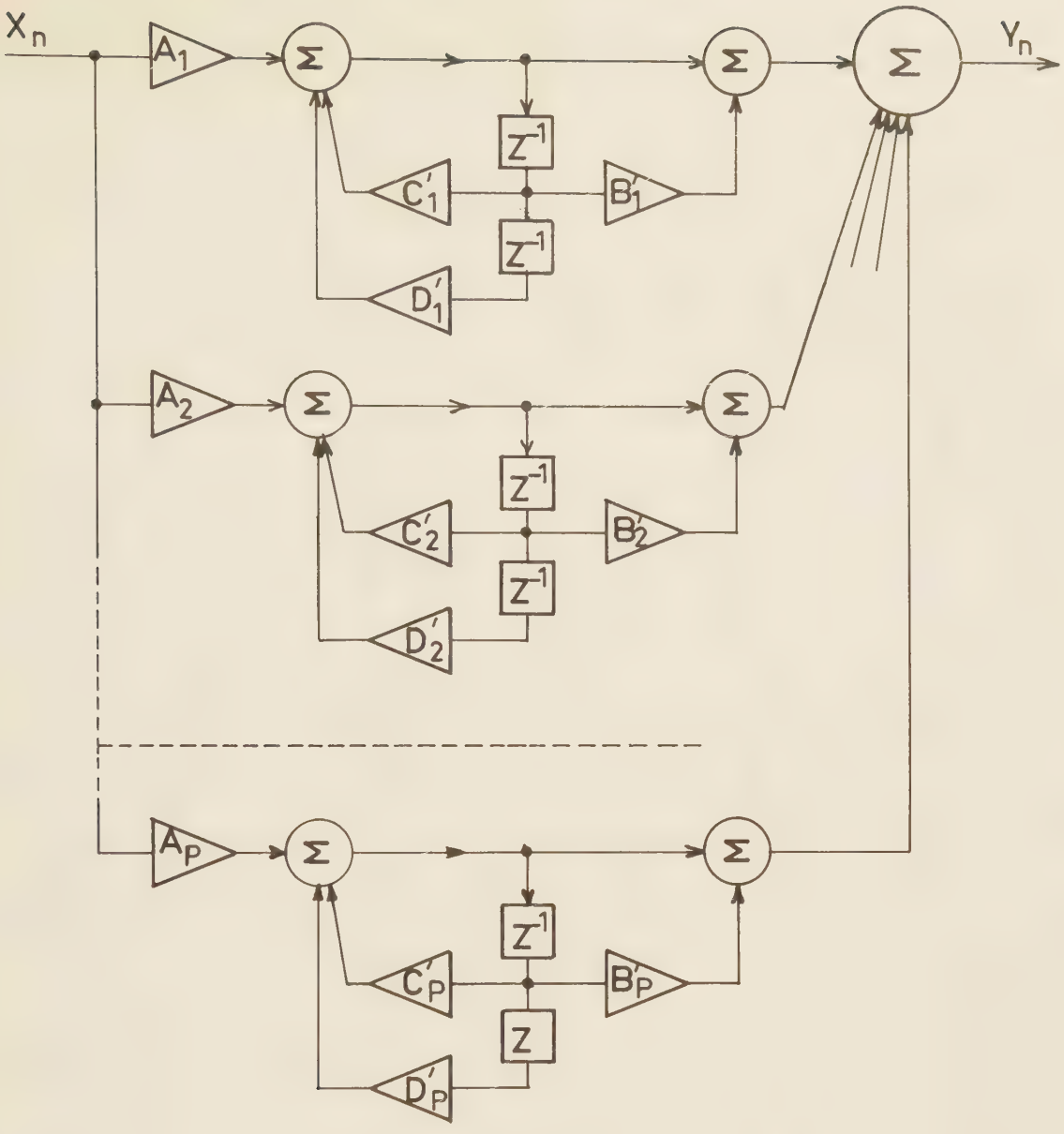


Figure 5.7

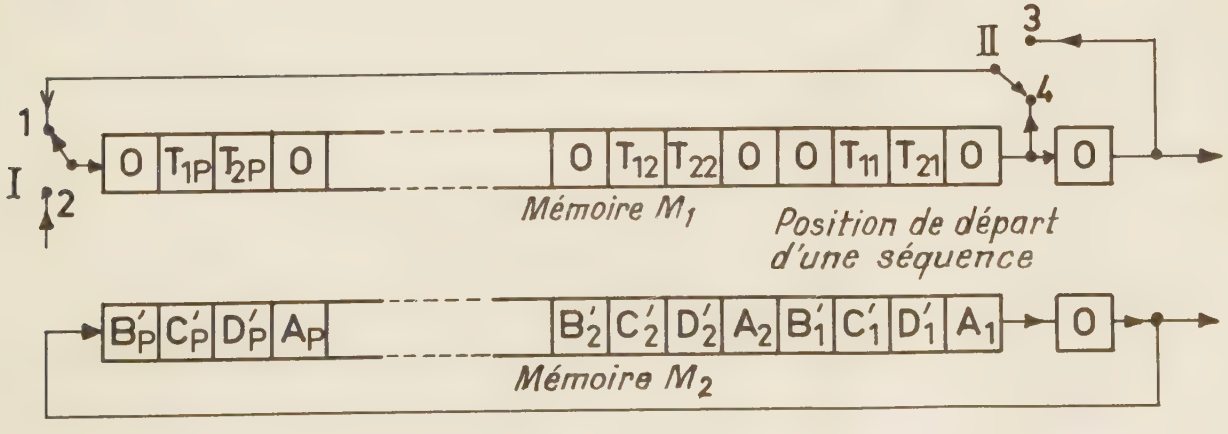


Figure 5.9

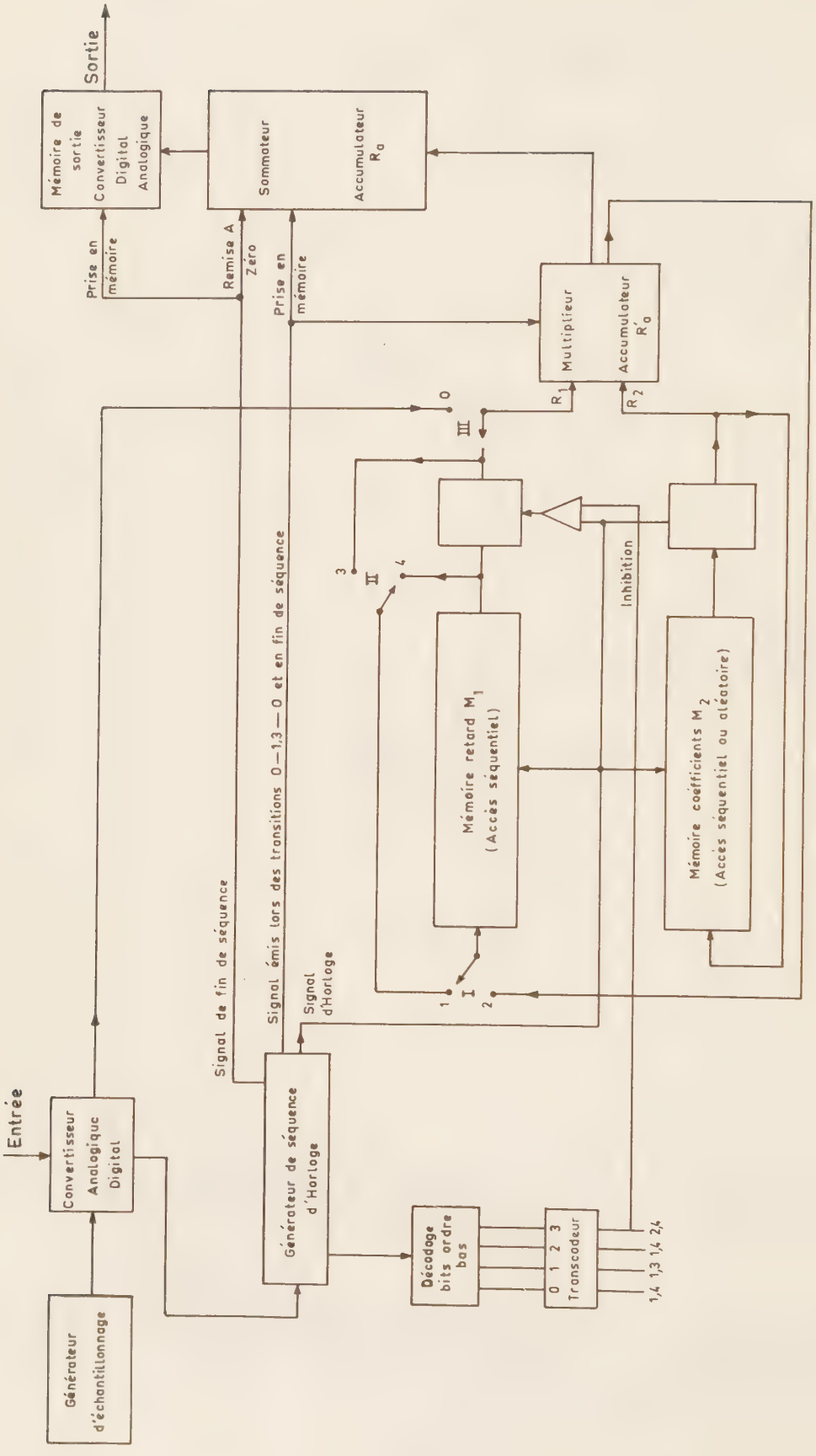


Figure 5.8 - Schéma synoptique d'un filtre numérique récursif programmable.

Chapitre 6

TESTS D'UTILISATION D'UN FILTRE NUMERIQUE RECURSIF PROGRAMMABLE DANS UN PROBLEME DE FILTRAGE ADAPTE

A titre d'exemple, nous allons calculer les valeurs des coefficients correspondant à un filtre adapté, destiné à la détection d'un signal certain apparaissant à un instant aléatoire, dans un bruit blanc qui le masque.

I - CALCUL DES COEFFICIENTS DU FILTRE ADAPTE

Soit à calculer les coefficients du filtre adapté au signal $S(t) = e^{-\pi(t+1,5)^2}$ (impulsion gaussienne).

On sait, d'après la théorie du filtrage adapté, que ce filtre devra avoir pour réponse impulsionnelle : $H(t) = S^*(-t)$, soit : $H(t) = e^{-\pi(t-1,5)^2}$ (impulsion gaussienne centrée au point $t = 1,5$).

Le gain complexe d'un tel filtre, transformée de Fourier de $H(t)$ s'écrit : $h(\nu) = e^{-\pi\nu^2}.e^{-3i\pi\nu}$ pour $\nu = \pm 1,5$, $|h(\nu)| = 0,00083$ c'est-à-dire que le module du gain est réduit à moins du millième de sa valeur maximale, nous le considérons comme nul ce qui conduit à une fréquence de Shannon $\nu_S = 3$, donc une période d'échantillonnage maximale de $T_S = \frac{1}{3}$. Nous choisirons finalement pour période d'échantillonnage $T_e = 0,1$ ce qui sera largement suffisant pour éviter toute erreur notable due au repliment du spectre.

Enfin pour $t < 0$ et $t > 3$, $|H(t)| < 0,00083$ nous considérerons donc $H(t)$ nul pour $t < 0$ (filtre causal) et $t > 3$. La réponse impulsionnelle du filtre cherché sera donc définie par l'échantillonnage suivant :

$$G_n = e^{-\pi T_e^2 (n-15)^2} \qquad \begin{cases} T_e = 0,1 \\ 0 \leqslant n < 30 \end{cases}$$

$$G_n = 0 \qquad n < 0, \quad n \geqslant 30$$

Le calcul effectué selon la méthode du chapitre 3, à partir de la réponse impulsionnelle, définie par ses 30 premiers points donne les résultats suivants :

a_o = 1,000 000
a_1 = - 2,798 153
a_2 = 5,199 644
a_3 = - 3,981 000
a_4 = 3,829 534
a_5 = - 0,622 286
a_6 = 5,031 079
a_7 = - 2,221 335
a_8 = 6,910 490
a_9 = - 0,933 252
a_{10} = 3,588 585
a_{11} = 0,681 409
a_{12} = 0,921 624
a_{13} = 0,311 622
a_{14} = 0,084 705

b_o = 1,000 000
b_1 = - 5,285 111
b_2 = 12,535 188
b_3 = - 17,197 159
b_4 = 14,878 999
b_5 = - 9,505 044
b_6 = 8,459 898
b_7 = - 12,585 628
b_8 = 16,149 323
b_9 = - 14,947 104
b_{10} = 10,131 851
b_{11} = - 5,099 322
b_{12} = 1,856 053
b_{13} = - 0,442 463
b_{14} = 0,051 976

II - TEST DU FILTRE OBTENU

On superpose au signal S(t) défini en 30 points ainsi que nous l'avons exposé au paragraphe précédent un bruit blanc défini en 1 000 points par un sous programme Fortran existant en bibliothèque. Sur le calculateur, on filtre la séquence des 1 000 points par le filtre ayant les coefficients précédemment obtenus ; on peut alors constater sur le signal de sortie un maximum dont on vérifie qu'il correspond bien à la position du signal S(t) dans la séquence de bruit blanc.

CONCLUSION

Le filtrage numérique récursif auquel a été consacré l'essentiel de cette étude s'avère particulièrement bien adapté aux techniques digitales ainsi que le montre l'apparition récente sur le marché d'éléments de filtres numériques intégrés en technologie MOS/LSI.

Par ailleurs, la diffusion prochaine de ces circuits intégrés permettra, en les simplifiant considérablement, de faire passer au second plan les problèmes de réalisation technique des filtres numériques, et de se consacrer essentiellement au problème de leur emploi judicieux afin d'en exploiter toutes les possibilités.

C'est ici qu'intervient la délicate question de la détermination des coefficients, nécessitant à la fois une bonne connaissance théorique générale du filtrage numérique et une adaptation au problème de filtrage envisagé.

Nous espérons que l'approche de ce problème dans le domaine temporel, convenant particulièrement au cas de la synthèse des filtres adaptés et que nous avons choisi d'étudier dans ce mémoire aura contribué à développer l'emploi du filtrage numérique dans un domaine où il est encore peu usité.

ANNEXE I

CALCUL DE LA FONCTION DE GREEN D'UN FILTRE NUMERIQUE RECURSIF

Il s'agit de calculer la solution de l'équation (16) :

$$\sum_{n=o}^{N} b_n . G_{k-n} = \delta_{k,o}$$

avec des conditions initiales homogènes.

Les solutions de l'équation homogène associée :

$$\sum_{n=o}^{N} b_n . G_{k-n} = 0$$

s'écrivent $V_i(k) = k^n . z_i^k$ $(n=0,1,\ldots,p-1)$, z_i étant une racine d'ordre p de l'équation algébrique :

$$\sum_{n=o}^{N} b_n . z^{N-n} = 0 \qquad (18)$$

Cette équation possède ainsi N racines $V_i(k)$ $(i=1,2,\ldots,N)$ linéairement indépendantes.

Cherchons pour l'équation (16) une solution de la forme :

$$G(k) = \sum_{i=1}^{N} U_i(k) . V_i(k)$$

les $U_i(k)$ étant des fonctions de k à déterminer. En utilisant la notation du produit scalaire de deux vecteurs U(k) et V(k) de composantes $U_i(k)$ et $V_i(k)$ dans un espace à N dimensions,

nous écrirons : $<G(k) = <U(k)|V(k)>$ (16a)

nous poserons également : $\Delta U(k) = U(k-1) - U(k)$ (16b)

Ecrivons alors :

$$G(k-1) = <V(k-1)|V(k-1)> = <U(k)|V(k-1)> + <\Delta U(k)|V(k-1)>$$

soit : $G(k-1) = <U(k)|V(k-1)>$ avec la condition $<\Delta U(k)|V(k-1)> = 0$

de même : $G(k-2) = <U(k)|V(k-2)>$ avec la condition $<\Delta U(k)|V(k-2)> = 0$

etc... $G(k-N+1) = <U(k)|V(k-N+1)>$ avec la condition $<\Delta U(k)|V(k-N+1)> = 0$

$$G(k-N) = <U(k)|V(k-N)> + <\Delta U(k)|V(k-N)>$$

Portant ces valeurs de G(k-n) dans l'équation (16), il vient, compte-tenu des conditions sur $\Delta U(k)$:

$$\langle \Delta U(k) \,|\, V(k-N) \rangle = \frac{\delta_{k,o}}{b_N}$$

Cette équation jointe aux N-1 conditions sur $\Delta U(k)$ forment un système de N équations linéaires permettant de déterminer les N composantes $\Delta U_i(k)$ de $\Delta U(k)$. Le déterminant de ce système s'écrit :

$$W(k-1) = \begin{vmatrix} V_1(k-1) & \cdots & V_N(k-1) \\ \vdots & & \vdots \\ V_1(k-N) & \cdots & V_N(k-N) \end{vmatrix}$$

Il n'est pas nul puisque les solutions $V_i(k)$ de l'équation homogène sont indépendantes. Nous obtenons donc :

$$\Delta U_i(k) = \frac{W_{N,i}(k-1)}{W(k-1)} \cdot \frac{\delta_{k,o}}{b_N} \qquad (16c)$$

$W_{N,i}(k)$ désignant le cofacteur de l'élément $V_i(k-N)$ dans $W(k-1)$.

Les conditions initiales homogènes se traduisent par $U(o) = 0$, on peut donc écrire :

$$U(k) = U(k)-U(k-1)+U(k-1)-U(k-2)+\ldots.-U(1)+U(1)-U(o)$$

d'où $$U(k) = -\sum_{j=o}^{k} \Delta U(j)$$

soit, d'après (16c) : $$U_i(k) = \sum_{j=o}^{k} \frac{W_{N,i}(j-1)}{W(j-1)} \cdot \frac{\delta_{j,o}}{b_N}$$

c'est-à-dire :
$$\begin{cases} k \geqslant 0 & U_i(k) = \dfrac{W_{N,i}(-1)}{W(-1).b_N} \\ k < 0 & U_i(k) = 0 \end{cases}$$

d'où finalement la solution :

$$\begin{cases} k \geqslant 0 & G_k = \displaystyle\sum_{n=1}^{N} \frac{W_{N,n}(-1)}{b_N.W(-1)} \cdot V_n(k) \\ k < 0 & K_k = 0 \end{cases}$$

ANNEXE II

ALGORITHME DE CALCUL D'UN SPECTRE - ETUDE DE L'ERREUR D'ESTIMATION

I - ALGORITHME DE CALCUL

On calculera le spectre ou densité spectrale énergétique (d.s.e.) d'un signal X(t) par application du théorème de Wiener Kinchine c'est-à-dire en calculant la fonction de corrélation puis sa transformée de Fourier. Le calcul étant fait sur ordinateur on calculera d'abord la fonction de corrélation en utilisant l'expression :

$$C_K(nT_e) = \sum_{k=-K}^{K} X(kT_e).X\left[(k-n).T_e\right] \qquad (1)$$

T_e désignant la période d'échantillonnage du signal, $C_K(nT_e)$ est une estimation de la valeur $C(nT_e)$ de la fonction de corrélation $C(t)$ au point $t = nT_e$, l'erreur d'estimation étant due au fait que K est nécessairement fini.

Nous supposerons de plus que le signal X(t) est à densité spectrale énergétique bornée, ou plus précisément que sa fonction de corrélation C(t) appartient à l'espace BL^2. Dans ces conditions la connaissance des échantillons $C(nT_e)$ suffit au calcul de la d.s.e. S(ν) du signal, en effet nous avons :

$$C(t) = \sum_{n=-\infty}^{+\infty} C(nT_e).\mathrm{sinc}\ \frac{\pi}{T_e}(t-nT_e) \qquad (2)$$

d'où par transformation de Fourier :

$$S(\nu) = T_e.\ \Pi_{\frac{1}{2T_e}}(\nu).\ \sum_{n=-\infty}^{+\infty} C(nT_e).e^{-2i\pi\nu nT_e} \qquad (3)$$

C(t) étant l'autocorrélation d'une fonction réelle paire, sa transformée de Fourier S(ν) est aussi une fonction réelle paire et l'on peut écrire :

$$S(\nu) = T_e.C(o) + 2T_e.\ \Pi_{\frac{1}{2T_e}}(\nu).\ \sum_{n=1} C(nT_e).\cos(2\pi\nu nT_e) \qquad (4)$$

En fait dans le cas pratique la sommation est nécessairement finie et de plus on ne connaît pas $C(nT_e)$ mais seulement $C_K(nT_e)$, on calculera donc en fait l'estimation $S_{N,K}(\nu)$ de S(ν) donnée par :

$$S_{N,K}(\nu) = T_e.C_K(o) + 2T_e.\ \Pi_{\frac{1}{2T_e}}(\nu).\ \sum_{n=1}^{N} C_K(nT_e).\cos\ (2\pi\nu nT_e) \qquad (5)$$

Enfin, le spectre $S(\nu)$ s'étendant de $-\frac{1}{2T_e}$ à $\frac{1}{2T_e}$ on en calculera 2P+1 points correspondant aux valeurs $\nu = \pm \frac{p}{p} \frac{1}{2T_e}$ avec $p = 0, 1, \ldots, P-1$, on aura donc finalement :

$$S_{N,K}\left(\frac{p}{2T_e.P}\right) = T_e.C_K + 2T_e \cdot \sum_{n=1}^{N} C_K(nT_e) \cos\left(\frac{\pi}{P}.n.p\right) \qquad (6)$$

II - ERREUR D'ESTIMATION SUR LE SPECTRE [8]

Le spectre calculé par (6) est finalement entaché de 2 erreurs d'estimation :

1° - Erreur due à la sommation finie dans le calcul de $C(nT_e)$ qui nous oblige dans l'équation (6) à employer les coefficients $C_K(nT_e)$ seuls connus, au lieu des coefficients exacts $C(nT_e)$.

On réduit cette erreur en augmentant le nombre K d'échantillons du signal.

2° - Erreur due à la sommation finie dans le calcul de $S(\nu)$. On montre que cette erreur d'estimation diminue lorsqu'on diminue T_e (échantillonnage plus fin) et que l'on augmente le temps $(2N+1)T_e$ sur lequel la fonction de corrélation est connue, ce qui globalement revient à augmenter le nombre N de points de corrélation.

En résumé on a :

$$S(\nu) = \lim_{\substack{K\to\infty \\ N\to\infty}} S_{N,K}(\nu)$$

ANNEXE III

CALCUL DES MOMENTS DU 1er ET 2ème ORDRE D'UNE VARIABLE ALEATOIRE QUANTIFIEE ET DU MOMENT MIXTE DE DEUX VARIABLES ALEATOIRES QUANTIFIEES

I - MOMENTS D'UNE VARIABLE

Soit X une variable aléatoire continue. Nous appelerons transformée de X par quantification la variable aléatoire discrète $\hat{X}$ définie par :

$$\hat{X} = n.q \qquad \text{si} \qquad X \in \left| (n-\tfrac{1}{2})q,\ (n+\tfrac{1}{2})q \right|$$

$n \in Z$; $q \in R$ (pas de quantification)

Soit $\phi(u)$ la fonction caractéristique de X et $\hat{\phi}(u)$ celle de $\hat{X}$ on démontre [4] la relation suivante :

$$\phi(u) = \sum_{k=-\infty}^{+\infty} \phi\left(u+\frac{2k\pi}{q}\right) \cdot \sin \frac{q}{2}\left(u+\frac{2k\pi}{q}\right) \qquad (1)$$

1° - Nous calculerons $E\{X\}$ par la formule $E\{X\} = \frac{1}{i}\left|\frac{d\phi(u)}{du}\right|_{u=o}$

D'après (1) il vient en posant $G(u) = \text{sinc}\ \frac{q.u}{2}$

$$\frac{d\hat{\phi}(u)}{du} = \sum_{k=-\infty}^{+\infty} \frac{d\phi\left(u+\frac{2k\pi}{q}\right)}{du} \cdot G\left(u+\frac{2k\pi}{q}\right) + \sum_{k=-\infty}^{+\infty} \phi\left(u+\frac{2k\pi}{q}\right) \frac{dG\left(u+\frac{2k\pi}{q}\right)}{du} \qquad (3)$$

- le 1er terme pris pour $u = o$ s'écrit :

$$\left[\frac{d\phi}{du}\right]_{u=o} = i.E\{X\} \qquad (4)$$

$\left(\text{car } G\left(\frac{2k\pi}{q}\right) = 0 \text{ pour } k \neq 0 \text{ et } G(o) = 1\right)$.

- le 2ème terme nécessite le calcul de $\left[\frac{dG\left(u+\frac{2k\pi}{q}\right)}{du}\right]_{u=o}$ qui donne pour résultat $(i1)^k \frac{q}{2k}$ pour $k \neq 0$ et 0 pour k=0, d'où finalement compte-tenu des relations (2), (3) et (4) :

$$E\{\hat{X}\} = E\{X\} + \frac{1}{i}\frac{q}{2\pi} \sum_{\substack{k=-\infty \\ k\neq 0}}^{+\infty} \frac{(-1)^k}{k} \phi\left(\frac{2k\pi}{q}\right)$$

C'est-à-dire, compte-tenu de la symétrie hermitienne de la fonction caractéristique :

$$E\{\hat{X}\} = E\{X\} + \frac{q}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \operatorname{Im}\left[\phi\left(\frac{2k\pi}{q}\right)\right]$$

2° - Le moment du 2ème ordre s'obtient de la même manière à partir de la formule

$$E\{\hat{X}^2\} = -\left[\frac{d^2\,\hat{\phi}(u)}{du^2}\right]_{u=o} \qquad (5)$$

En dérivant (3), il vient :

$$\frac{d^2\,\hat{\phi}(u)}{du^2} = \sum_{k=-\infty}^{+\infty} \frac{d^2}{du^2}\left[\phi\left(u+\frac{2k\pi}{q}\right)\right].G(u) + \sum_{k=-\infty}^{+\infty} \phi\left(u+\frac{2k\pi}{q}\right).\frac{d^2G\left(u+\frac{2k\pi}{q}\right)}{du^2}$$

$$+ 2\sum_{k=-\infty}^{+\infty} \frac{d\phi\left(u+\frac{2k\pi}{q}\right)}{du}.\frac{dG\left(u+\frac{2k\pi}{q}\right)}{du}$$

- le 1er terme donne pour u = 0 $\left[\frac{d^2\,\phi(u)}{du^2}\right]_{u=o} = = -E\{X^2\}$

- le 2ème terme nécessite le calcul de $\left[\frac{d^2G\left(u+\frac{2k\pi}{q}\right)}{du^2}\right]_{u=o}$ qui donne :

$-\frac{q^2}{2\pi^2}\frac{(-1)^k}{k^2}$ pour $k\neq 0$ et $-\frac{q^2}{12}$ pour $k=0$

- le 3ème terme enfin utilise le résultat déjà obtenu :

$$\left[\frac{dG\left(u+\frac{2k\pi}{q}\right)}{du}\right]_{u=o} = (-1)^k\,\frac{q}{2k\pi}$$

Il vient alors compte-tenu de (5) :

$$E\{X^2\} = E\{X^2\} + \frac{q^2}{12}+\frac{q^2}{2\pi^2}.\sum_{\substack{k=-\infty \\ k\neq o}}^{+\infty} \frac{(-1)^k}{k^2}.\phi\left(\frac{2k\pi}{q}\right) - \frac{q}{\pi}.\sum_{\substack{k=-\infty \\ k\neq o}}^{+\infty}.\phi'\left(\frac{2k\pi}{q}\right)$$

Compte-tenu de la symétrie hermitique de $\phi(u)$: $\phi(-u)$ $\phi^*(u)$ qui entraîne aussi $-\phi'(-u) = \phi'^*(u)$ il vient finalement la relation :

$$E\{\hat{X}^2\} = E\{X^2\} + \frac{q^2}{12} + \frac{q^2}{\pi^2} \cdot \sum_{k=1}^{\infty} \frac{(-1)^k}{k^2} \cdot \mathcal{R}_e \left[\phi\left(\frac{2k\pi}{q}\right) \right] - \frac{2q}{\pi} \cdot \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \cdot \mathcal{R}_e \left[\phi'\left(\frac{2k\pi}{q}\right) \right]$$

II - MOMENT DE DEUX VARIABLES

Soit $X(X_1,X_2)$ une variable aléatoire continue à deux dimensions. Nous appellerons transformée de X par quantification la variable aléatoire discrète à deux dimensions $\hat{X}(\hat{X}_1,\hat{X}_2)$ définie par :

$$\hat{X}_1 = n.q \quad \text{si} \quad X_1 \in \left[\left(n-\tfrac{1}{2}\right) q, \left(n+\tfrac{1}{2}\right) q \right]$$

$$\hat{X}_2 = m.q \quad \text{si} \quad X_2 \in \left[\left(m-\tfrac{1}{2}\right) q, \left(m+\tfrac{1}{2}\right) q \right]$$

$m, n \in Z$; $q \in R$ (q = pas de quantification)

Soit $\phi(u_1,u_2)$ la fonction caractéristique de X et $\hat{\phi}(u_1,u_2)$ celle de $\hat{X}$ on obtient la relation suivante par analogie avec la relation (1) :

$$\hat{\phi}(u_1,u_2) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} \phi\left(u_1+\frac{2k_1\pi}{q}, u_2+\frac{2k_2\pi}{q}\right) \cdot \operatorname{sinc}\frac{q}{2}\left(u_1+\frac{2k_1\pi}{q}\right) \cdot \operatorname{sinc}\frac{q}{2}\left(u_2+\frac{2k_2\pi}{q}\right) \qquad (6)$$

Nous calculerons $E\{\hat{X}_1.\hat{X}_2\}$ par la formule :

$$E\{\hat{X}_1.\hat{X}_2\} = - \frac{\partial \hat{\phi}}{\partial u_1 . \partial u_2}$$

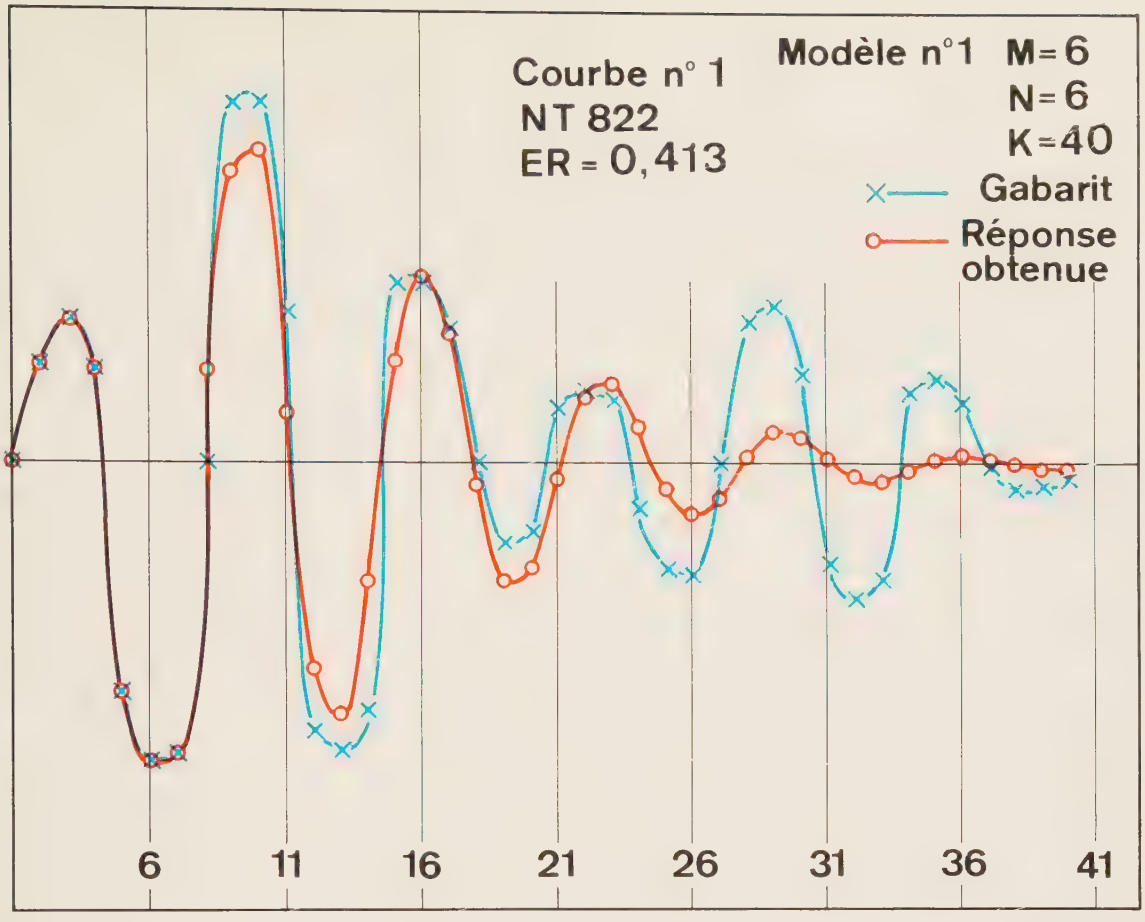
Un calcul analogue à celui du paragraphe précédent nous donne :

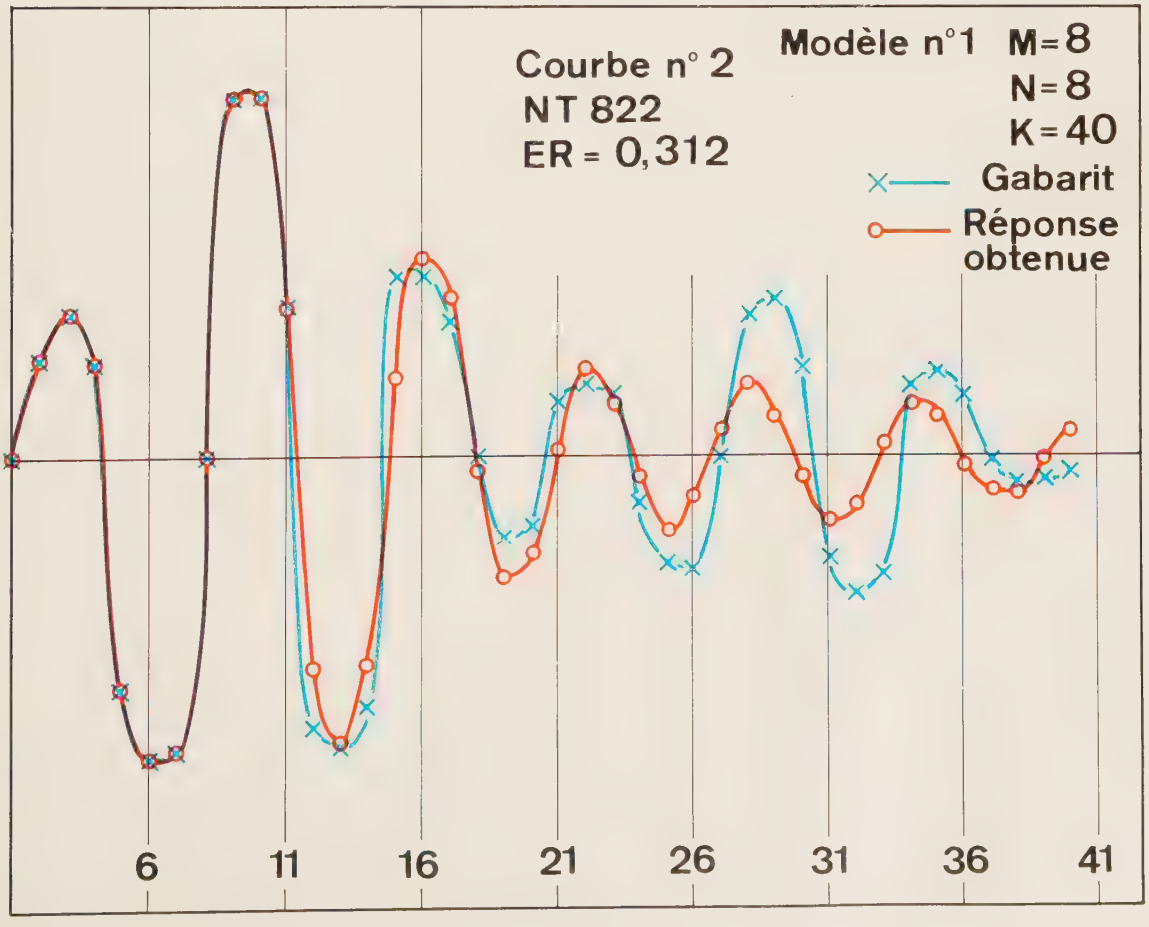
$$E\{\hat{X}_1.\hat{X}_2\} = E\{X_1.X_2\} + \frac{q^2}{2\pi^2} \sum_{\substack{k_1=-\infty \\ k_1\neq 0}}^{+\infty} \sum_{\substack{k_2=-\infty \\ k_2\neq 0}}^{+\infty} \frac{(-1)^{k_1+k_2}}{k_1.k_2} \cdot \phi\left(\frac{2k_1\pi}{q}, \frac{2k_2\pi}{q}\right)$$

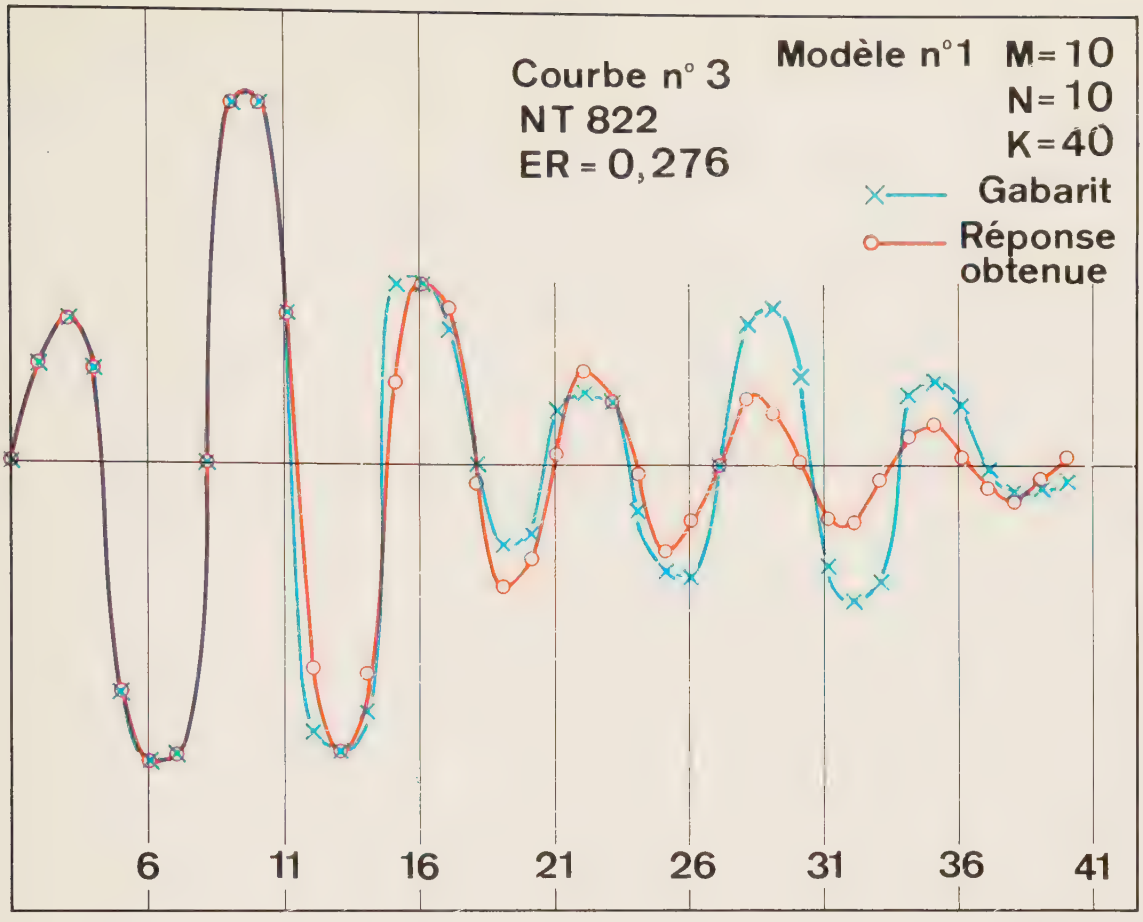
$$- \frac{q}{2\pi} \sum_{k_1=-\infty}^{+\infty} \frac{(-1)^{k_1}}{k_1} \frac{\partial \phi}{\partial u_2} \left(\frac{2k_1\pi}{q}, 0\right) - \frac{q}{2\pi} \sum_{\substack{k_2=-\infty \\ k_2\neq 0}}^{+\infty} \frac{(-1)^{k_2}}{k_2} \cdot \frac{\partial \phi}{\partial u_1} \left(0, \frac{2k_2\pi}{q}\right)$$

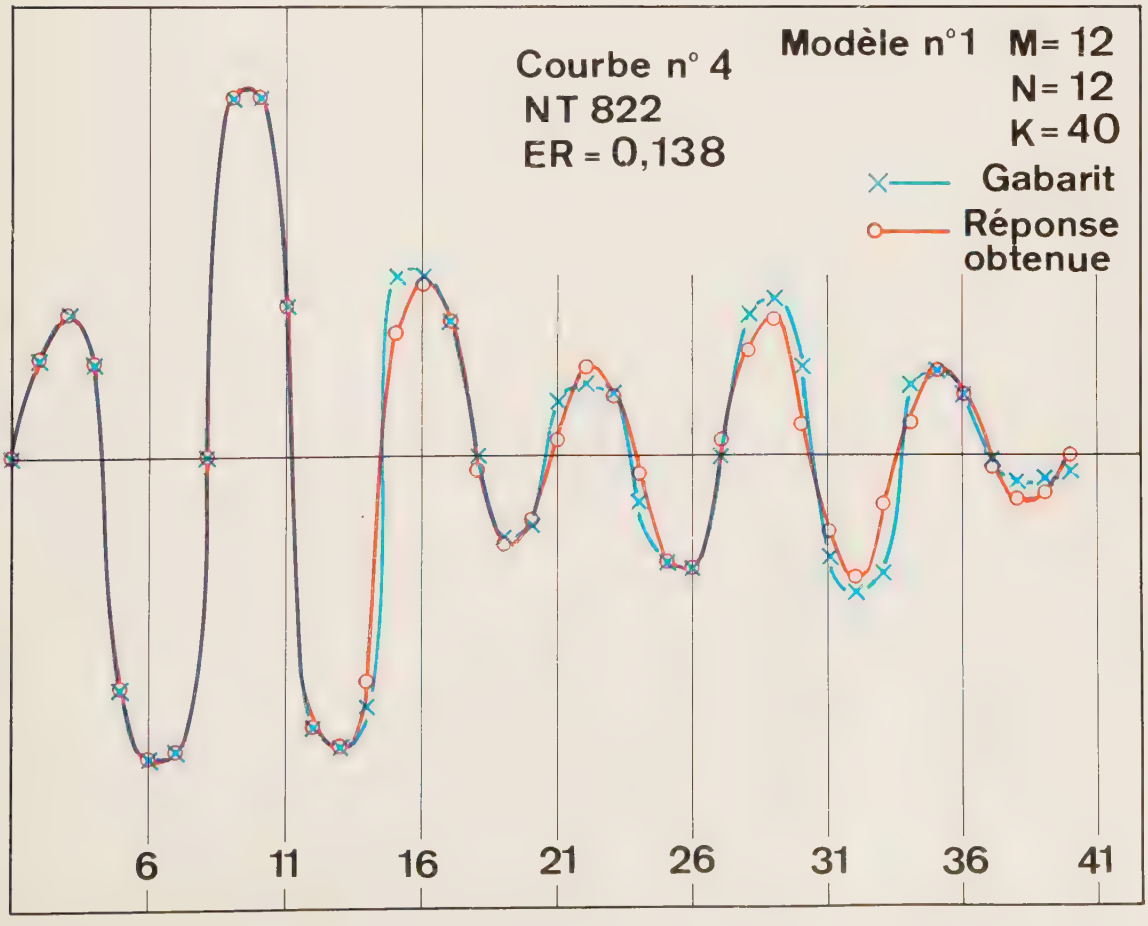
ANNEXE IV

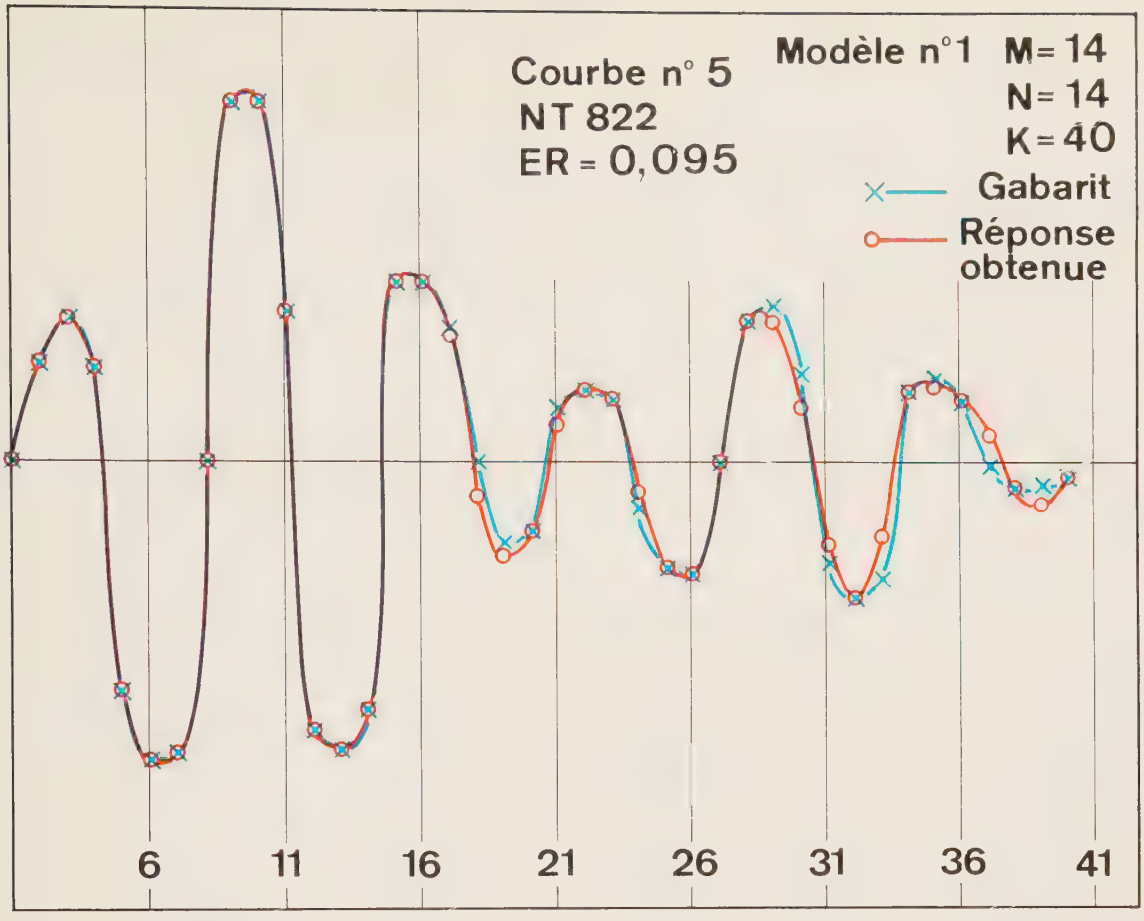
APPROCHE D'UN MODELE DE FILTRE DONNE DANS LE DOMAINE TEMPOREL PAR DES FILTRES NUMERIQUES RECURSIFS D'ORDRE CROISSANT

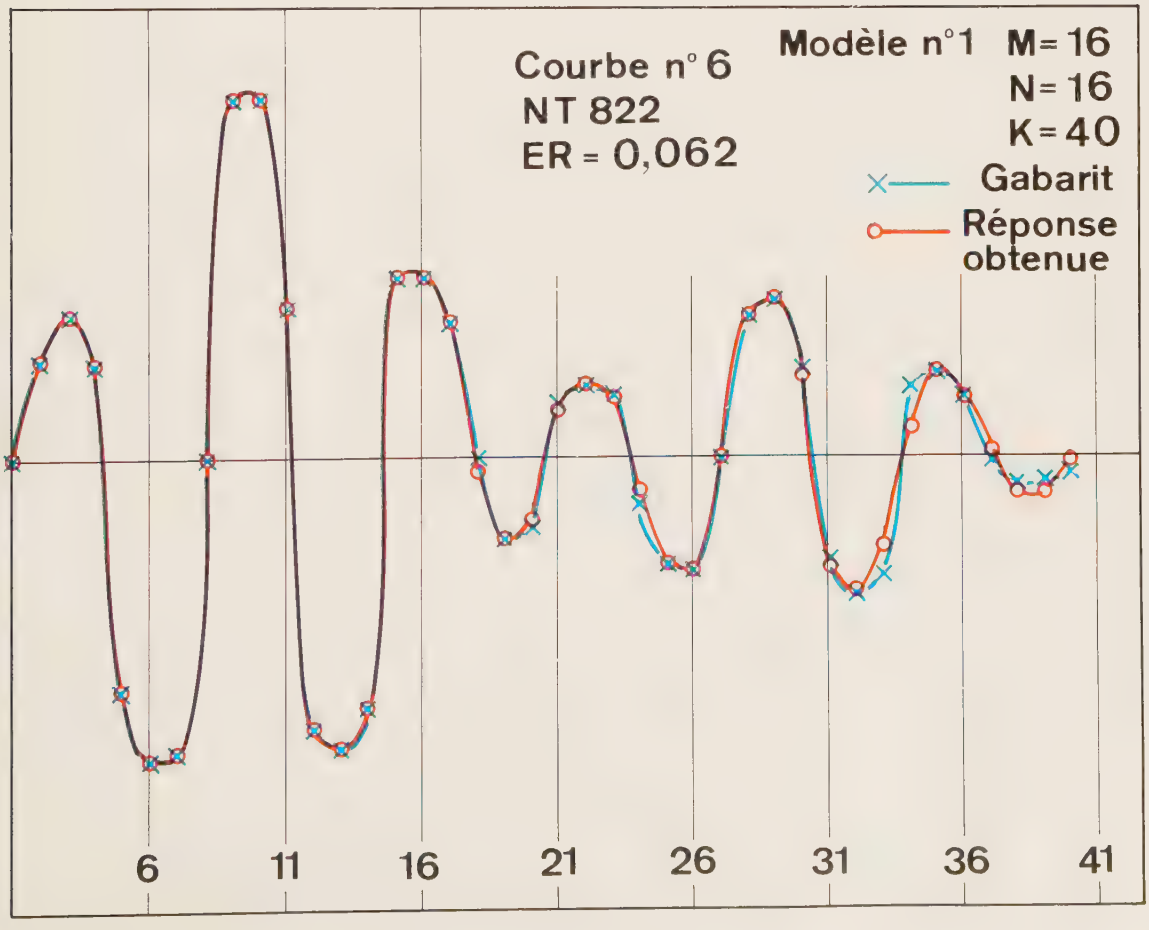

Courbe n° 1
NT 822
ER = 0,413
Modèle n°1 M=6
N=6
K=40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41

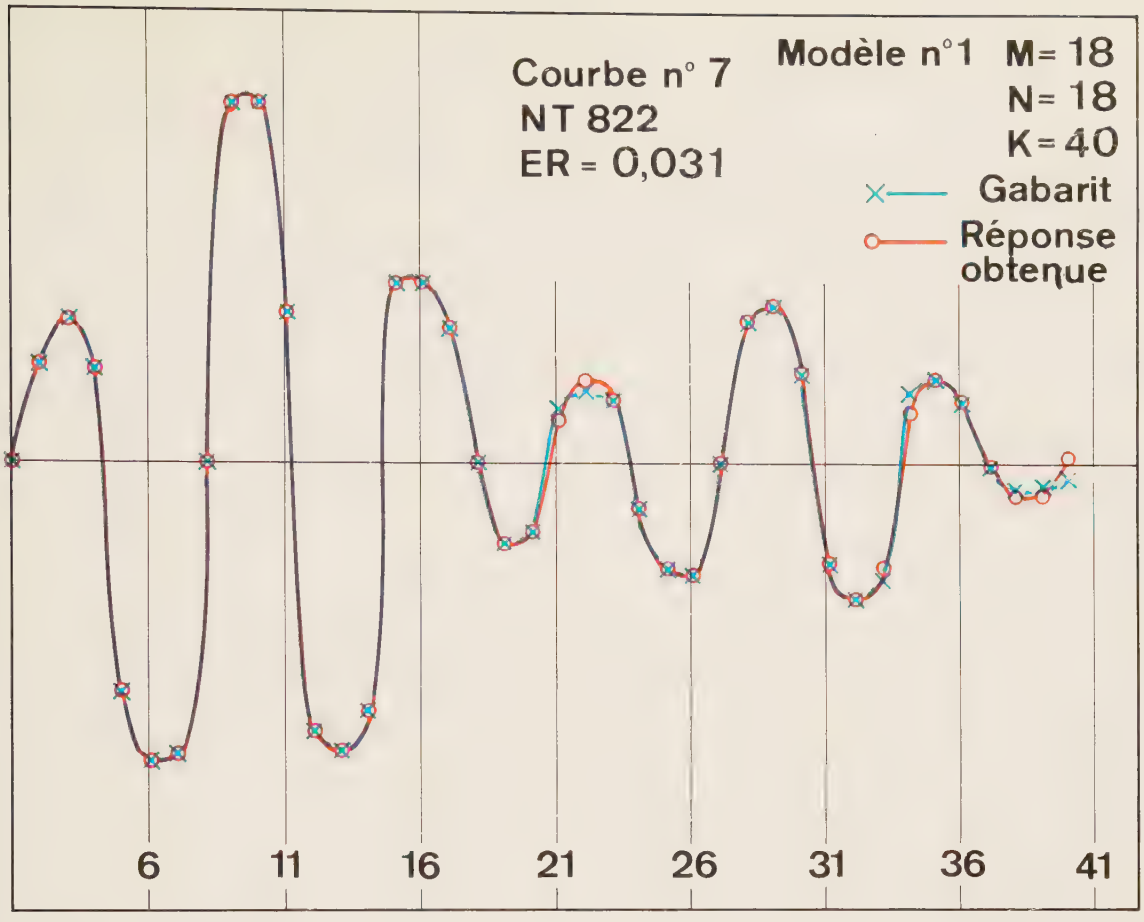

Courbe n° 2
NT 822
ER = 0,312
Modèle n°1 M=8
N=8
K=40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41

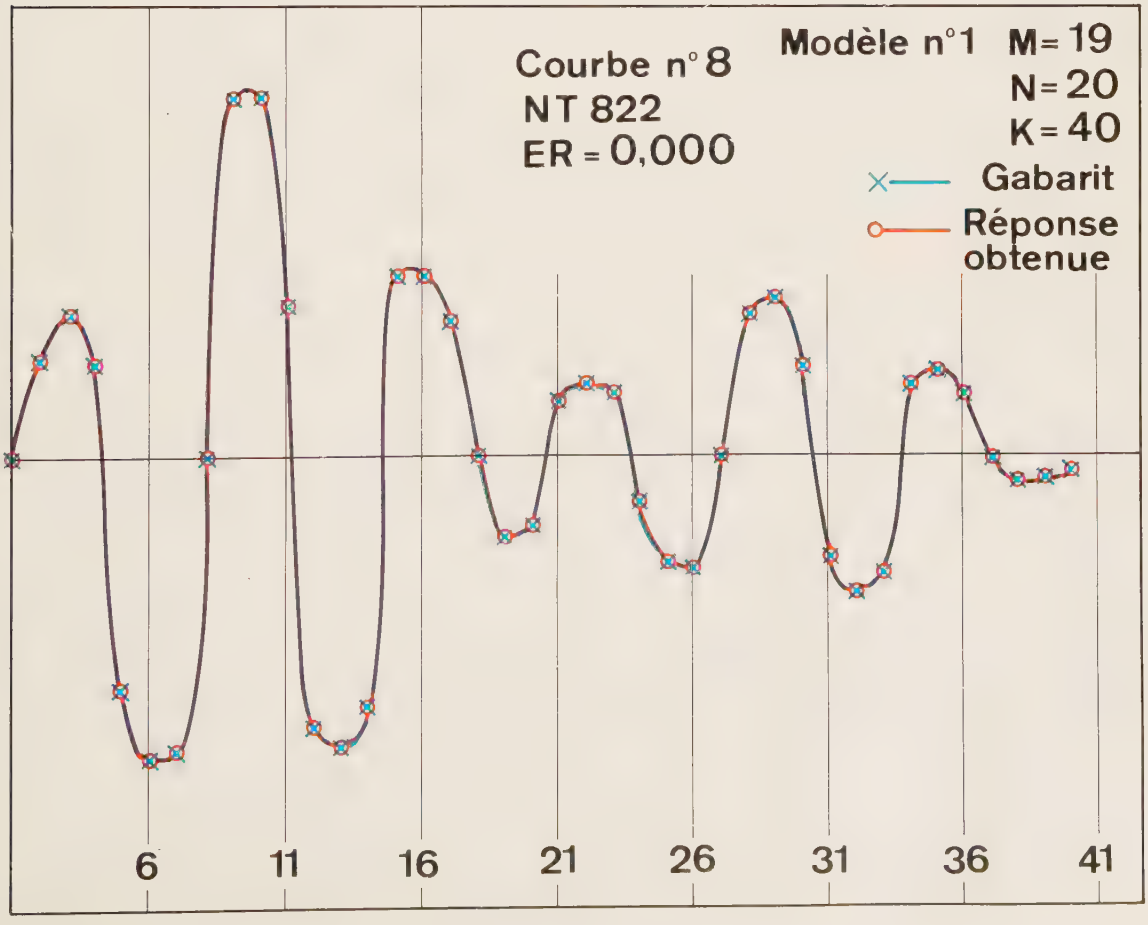

Courbe n° 3
NT 822
ER = 0,276
Modèle n°1 M=10
N=10
K=40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41


Courbe n° 4
NT 822
ER = 0,138
Modèle n°1 M=12
N=12
K=40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41


Courbe n° 5
NT 822
ER = 0,095
Modèle n°1 M= 14
N= 14
K= 40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41


Courbe n° 6
NT 822
ER = 0,062
Modèle n°1 M= 16
N= 16
K= 40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41


Courbe n° 7
NT 822
ER = 0,031
Modèle n°1 M= 18
N= 18
K= 40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41


Courbe n° 8
NT 822
ER = 0,000
Modèle n°1 M= 19
N= 20
K= 40
Gabarit
Réponse obtenue
6
11
16
21
26
31
36
41

BIBLIOGRAPHIE

[1] BONNET G.
Communication présentée à la Conférence sur le Traitement du Signal à Paris le 7 mars 1962.
Note technique CEA.CENG EL/BS n° 3

[2] BURRUS C.S., PARKS T.W.
Time domain design of recursive digital filters IEEE Transaction on Audio and Electroacousties.
Vol. AU.18 n° 2, juin 1970

[3] CHAUDOURNE S.
Synthèse des filtres numériques récursifs dans le domaine temporel.
Note technique CENG LETI/MCTE n° 828, avril 1972

[4] CHAUDOURNE S., KOFMAN W.
Etude d'un corrélateur utilisant deux sources de bruits auxiliaires.
Note technique CENG LETI/MCTE n° 725, mai 1971

[5] DURAND E.
Solutions numériques des équations algébriques, Tome II : systèmes de plusieurs équations.
Valeurs propres des matrices MASSON 1961

[6] EL MALAWANY I.
Filtrage numérique. Thèse de Doctorat d'Université Faculté des Sciences de l'Université de Rennes. Février 1968

[7] EL MALAWANY I.
Le filtrage numérique.
Annales des Télécommunications t.24, n° 3.4, 1969

[8] FAUQUE J.M.
Etude et réalisation d'un analyseur de densité spectrale énergétique.
Thèse d'Ingénieur Docteur.
Faculté des Sciences de l'Université de Grenoble, juillet 1970

[9] FAUQUE J.M.
Le problème de la déconvolution. Exposé d'une méthode de résolution matricielle. Note technique CENG LETI/MCTE n° 39 septembre 1969
Matrices du second degré et normes générales en analyse numérique linéaire.

[10] GASTINEL N.
Thèse de Docteur ès-sciences mathématiques.
Faculté des Sciences de l'Université de Grenoble, décembre 1960

[11] GOLD B., RADER C.M.
Digital processing of signals - MC GRAW HILL 1969

[12] GOLUB G.
Numerial Methods for solving linear least squares problems.
Numerische Mathematik 7, 206-216, 1965

[13] HUELSMAN
Active filters Mc Graw hill. chapitre 5 digital filters par John V. Wait

[14] KORGANOFF A., PAVEL M.
Par vu, méthodes de calcul numérique.
Tome II éléments de théorie des matrices carrées et rectangles. Dunod 1967

[15] MAX J.
Méthodes et techniques de Traitement du Signal.
Tome I : principes et appareillages de traitement en temps réel. Masson 1972

[16] RIESZ F., NAGY B. SZ
Leçons d'analyse fonctionnelle. Académie des sciences de Hongrie, 1955

[17] ROUBINE E.
Introduction à la théorie de la communication.
Tome I : Signaux non aléatoires
Tome II : Signaux aléatoires
Masson 1970

[18] SCHEIDECKER J.P.
Résolution d'une équation de convolution - Thèse de Doctorat du 3ème cycle.
Faculté des Sciences de Paris.

[19] SCHWARTZ L.
Méthodes mathématiques pour les sciences physiques, Hermann, 1965

[20] SMIRNOV V.I.
A course of Higher Mathematics. Volume V : Integration and functional analysis - Pergamon Press, 1964

Manuscrit reçu le 26 novembre 1973

VU

Grenoble, le 5 juin 1973

VU, et permis d'imprimer

Grenoble, le 29 juin 1973